Published in final edited form as:
Comput Stat Data Anal. 2008 March 15; 52(7): 3528–3542. doi:10.1016/j.csda.2007.11.007.

Efficient methods for estimating constrained parameters with applications to lasso logistic regression

Guo-Liang Tian[†,*], **Man-Lai Tang**[‡], **Hong-Bin Fang**[†], and **Ming Tan**[†]

†*Division of Biostatistics, University of Maryland Greenebaum Cancer Center, 10 South Pine Street, MSTF Suite 261, Baltimore, Maryland 21201, U.S.A.*

‡*Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, P. R. China*

Abstract

Fitting logistic regression models is challenging when their parameters are restricted. In this article, we first develop a *quadratic lower-bound* (QLB) algorithm for optimization with box or linear inequality constraints and derive the fastest QLB algorithm corresponding to the smallest global majorization matrix. The proposed QLB algorithm is particularly suited to problems to which EM-type algorithms are not applicable (e.g., logistic, multinomial logistic, and Cox's proportional hazards models) while it retains the same EM ascent property and thus assures the monotonic convergence. Secondly, we generalize the QLB algorithm to penalized problems in which the penalty functions may not be totally differentiable. The proposed method thus provides an alternative algorithm for estimation in lasso logistic regression, where the convergence of the existing lasso algorithm is not generally ensured. Finally, by relaxing the ascent requirement, convergence speed can be further accelerated. We introduce a pseudo-Newton method that retains the simplicity of the QLB algorithm and the fast convergence of the Newton method. Theoretical justification and numerical examples show that the pseudo-Newton method is up to 71 (in terms of CPU time) or 107 (in terms of number of iterations) times faster than the fastest QLB algorithm and thus makes bootstrap variance estimation feasible. Simulations and comparisons are performed and three real examples (Down syndrome data, kyphosis data, and colon microarray data) are analyzed to illustrate the proposed methods.

Keywords

Constrained parameter problems; EM algorithm; Lasso-type algorithm; Logistic regression; Pseudo-Newton algorithm; QLB algorithm; Variable selection

1. Introduction

Logistic regression is one of the most widely used statistical tools in many areas such as biomedicine, social sciences, economics and business (Collett, 1991; Agresti, 2002). If we know that parameters are restricted by some constraints, then it is reasonable to expect that we should be able to do better by incorporating such additional information than by ignoring them (Robertson *et al.*, 1988; Silvapulle and Sen, 2005). Fitting logistic models becomes challenging

*Corresponding author's email: gtian2@umm.edu.

when some model parameters are restricted inside a convex region in the Euclidean space (e.g., parameter estimation for lasso regression). When *maximum likelihood estimates* (MLEs) of parameters are located on the boundary of the region or the region that can be represented in terms of a set of equality/inequality restrictions, the constrained optimization problem may reduce to penalized problem that is closely related to the posterior mode (or maximum a *posteriori* estimate) in a Bayesian framework. The situation can be further complicated if the penalty function is not totally differentiable.

We consider the well-known lasso logistic regression which motivates the present problem of interest. Variable selection is one of the most pervasive problems in statistical applications. Classic methods for model/variable selection have not had much success in biomedical application, especially in high-dimensional data analysis including gene or protein expression data analysis, partly due to their numerical instability. A novel method that mitigates some of this instability and has good predictive performance is the lasso regression (Tibshirani, 1996). For logistic models, the lasso regression is to find

$$\widehat{\theta}^{\text{lasso}} = \arg\max \ell(\theta) \quad \text{subject to} \quad \sum_{j=1}^{q} |\theta_j| \le u, \tag{1.1}$$

where $\theta = (\theta_1, \ldots, \theta_q)^{\mathrm{T}}$ is a $q \times 1$ vector of unknown parameters, $\ell(\theta)$ is the log-likelihood function defined in (1.4) below, and u is a tuning parameter. Although a quadratic approximation to $\ell(\theta)$ can lead to a simpler iteratively reweighted least squares procedure (Tibshirani, 1996), convergence of this procedure is not generally ensured. One possible extension of (1.1) is to formulate the so-called bridge regression (Frank and Friedman, 1993). In this case, one would like to find

$$\widehat{\theta}^{\text{bridge}} = \arg\max \ell(\theta) \quad \text{subject to} \quad \sum_{j=1}^{q} |\theta_j|^{\gamma} \le u, \tag{1.2}$$

where $\gamma > 0$. However, solution to (1.2) was not given for any given u and γ in Frank and Friedman (1993).

Motivated by the constrained optimization problems (1.1) and (1.2), we consider the following logistic model with constrained parameters,

$$y_i \overset{\text{ind}}{\sim} \text{Binomial}(n_i, p_i), \quad \text{logit}(p_i) = x_{(i)}^{\mathrm{T}}\theta, \quad 1 \le i \le m, \tag{1.3}$$

where y_i denotes the number of subjects with positive response in n_i trials and $\{y_i\}_{i=1}^{m}$ are independent, p_i is the probability that a subject gives positive response, $x_{(i)}$ is the vector of covariates, and θ is a $q \times 1$ vector of unknown coefficients being restricted by some simple constraints $a \le \theta \le b$ for some $q \times 1$ constant vectors a and b (e.g., the lasso regression) or linear inequalities of the form $c \le P_{k\times q}\,\theta \le d$ for some $k \times 1$ constant vectors c and d (see the examples in §5.3 and §5.4). When $\text{rank}(P) = q$, letting $\mu = P\theta$ yields $a \le \mu \le b$ and $\theta = (P^{\mathrm{T}}P)^{-1}P^{\mathrm{T}}\mu$. In other words, linear inequality constraints with a full column-rank matrix P can be reparameterized into simple box constraints $[a, b]$ (Khuri, 1976;Tan *et al.*, 2003,2007). Hence, the log-likelihood function of θ in (1.3) is

$$\ell(\theta) = \sum_{i=1}^{m} \left\{ y_i (x_{(i)}^{\mathrm{T}}\theta) - n_i \log[1 + \exp(x_{(i)}^{\mathrm{T}}\theta)] \right\} \tag{1.4}$$

and the goal is to find the constrained MLE $\hat{\theta}$ or the penalized MLE $\tilde{\theta}$ given by

$$\widehat{\theta} = \arg\max_{\theta \in [a,b]} \ell(\theta), \quad \text{or} \tag{1.5}$$

$$\tilde{\theta}=\arg \max \{\ell(\theta) - \lambda J_1(\theta)\}, \tag{1.6}$$

where $J_1(\theta)$ is a penalty function and $\lambda > 0$ is a smoothing parameter for the tradeoff between the accuracy of the model fit and smoothness. When $J_1(\theta)$ is not totally differentiable, the penalized problem (1.6) sometimes can be reformulated as

$$\tilde{\theta}=\arg \max_{\theta\in[a,b]} \{\ell(\theta) - \lambda J_2(\theta)\}, \tag{1.7}$$

where $J_2(\theta)$ is differentiable everywhere.

When the log-likelihood is well-behaved (e.g., well approximated by a quadratic function), a natural algorithm for finding MLE is the Newton-Raphson (NR) or scoring methods because they converge quadratically. For logistic model with large number of variables (e.g., genes), both methods require tedious calculations for the Hessian or expected information matrix at each iteration. In addition, the log-likelihood does not necessarily increase at each iteration for NR method, which may sometimes be divergent (Cox and Oakes, 1984, p.172). Böhning and Lindsay (1988, p.645–646) provided an example of a concave function for which the NR method does not converge. Although EM-type algorithms (Dempster *et al.*, 1977; Meng and Rubin, 1993) possess the ascent property that ensures monotone convergence, they are not applicable to logistic regression owing to the absence of a missing-data structure. Therefore, for problems in which the missing-data structure does not exist or is not readily available, the *quadratic lower-bound* (QLB) algorithm (Böhning and Lindsay, 1988) is often an alternative. However, when the model has constrained parameters, the QLB algorithm is not applicable. In addition, like EM-type algorithms, the QLB algorithm is usually criticized for its slow convergence, especially for solving complicated problems or in high-dimensional data analysis.

In this paper, we first develop a QLB algorithm that can generally be applicable to optimization with box or linear inequality constraints. The fastest QLB corresponding to the smallest global majorization matrix is also derived. In brief, the QLB algorithm consists of an optimization transfer (T-step) and a constrained maximization (M-step). The T-step transfers the optimization from the intractable log-likelihood function to a quadratic surrogate function $Q(\theta|\theta')$ such that both functions share the same maximizer. The M-step can often be accomplished via some built-in SPLUS functions (e.g., nnls.fit or nlregb), which is one of the advantages of the algorithm. The QLB algorithm is especially suited to those problems (e.g., logistic, multinomial logistic, and Cox's model) in which EM-type algorithms are not applicable while it retains the same EM ascent property and thus assures the monotonic convergence. Secondly, we generalize the QLB algorithm to penalized problems in which the penalty functions may not be totally differentiable. The proposed method therefore provides an alternative algorithm for estimation in lasso logistic regression, where the convergence of the existing lasso algorithm is not generally ensured. Finally, to accelerate the convergence rate, we introduce a pseudo-Newton algorithm that does not necessarily have the ascent property but retains the simplicity of the QLB algorithm and the fast convergence of the Newton method. We show both theoretically and numerically that the pseudo-Newton algorithm is dramatically faster than the fastest QLB algorithm (up to 71 times in CPU time or 107 times in numbers of iterations) and thus makes the bootstrap variance estimation feasible. Another merit of the pseudo-Newton method is that the Cholesky decomposition of the surrogate matrix is calculated only once while the same matrix is required to be updated at each iteration for the Newton method.

The rest of this article is organized as follows. Section 2 develops the QLB algorithm for optimization with box or linear inequality constraints and derives the fastest QLB algorithm. Section 3 generalizes the QLB algorithm to penalized problems and investigates some convergence properties. Section 4 introduces a pseudo-Newton method. We apply the fastest

QLB and pseudo-Newton algorithm to estimate constrained parameters and to select variables in logistic regression in Section 5 and Section 6, respectively. Simulations and comparisons are performed and three published data sets are analyzed to illustrate the proposed methods. We conclude with a discussion in Section 7.

2. A QLB algorithm for optimization with box or linear inequality constraints

2.1 Formulation of the algorithm

Consider the calculation of the constrained MLE as follows

$$\widehat{\theta}=\arg\max_{\theta\in[a,b]}\ell(\theta), \tag{2.1}$$

where $\ell(\theta)$ is a twice continuously differentiable and concave function. Let ∇ denote the derivative operator, $\nabla\ell(\theta)$ the gradient vector and $\nabla^2\ell(\theta)$ the Hessian matrix. A key assumption for the QLB algorithm is that there exists a positive definite matrix B (denoted as $B > 0$) which globally majorizes the observed information, i.e.,

$$B \geq -\nabla^2\ell(\theta) \qquad \text{for all} \quad \theta \in \mathbb{R}^q, \tag{2.2}$$

where B does not depend on θ. Throughout this paper, a matrix B (> 0) is said to be a *global majorization* (GM) matrix if B is independent of θ and satisfies the condition (2.2). Furthermore, for two given θ and $\theta' \in \mathbb{R}^q$, a quadratic surrogate function is defined by

$$Q(\theta|\theta')=\ell(\theta')+(\theta-\theta')^{\mathrm{T}}\nabla\ell(\theta')-0.5(\theta-\theta')^{\mathrm{T}}B(\theta-\theta'). \tag{2.3}$$

Proposition 1 below shows that the QLB algorithm has the same EM ascent property. This property implies that finding (2.1) is equivalent to iteratively finding

$$\theta^{(t+1)}=\arg\max_{\theta\in[a,b]}Q(\theta|\theta^{(t)}) \tag{2.4}$$

provided that the initial value $\theta^{(0)} \in [a, b]$. To implement the M-step of the algorithm, we first construct an upper triangular matrix C via the Cholesky decomposition such that $B = C^{\mathrm{T}} C$, and define a q-vector depending on B and $\theta^{(t)}$ ($\theta^{(t)} \in [a, b]$) as

$$\xi=\xi(B,\theta^{(t)})=(C^{-1})^{\mathrm{T}}\nabla\ell(\theta^{(t)})+C\theta^{(t)}. \tag{2.5}$$

(2.4) thus becomes

$$\theta^{(t+1)}=\arg\min_{\theta\in[a,b]}\|\xi(B,\theta^{(t)})-C\theta\|^2. \tag{2.6}$$

Some built-in SPLUS functions such as nnls.fit (nonnegative least squares) and nlregb (nonlinear least squares subject to box constraints) can be utilized to calculate (2.6). Given a concave log-likelihood $\ell(\theta)$, the score vector $\nabla\ell(\theta)$ and the observed information matrix $-\nabla^2\ell(\theta)$, the QLB algorithm consists of

T-step: To find a GM matrix B and construct an upper triangular matrix C via the Cholesky decomposition such that $B = C^{\mathrm{T}} C$, and

M-step: To update the current estimate $\theta^{(t)}$ via (2.6)

2.2 The ascent property

One of the appealing features of the above QLB algorithm is that it possesses the EM ascent property. That is, the likelihood increases in each QLB iteration. We prove this ascent property in the following proposition.

Proposition 1—Under the assumptions (2.2) and (2.3), (i) $\ell(\theta) - Q(\theta|\theta')$ achieves its minimum (i.e., zero) at $\theta = \theta'$ for all $\theta, \theta' \in \mathbb{R}^q$; and (ii) the ascent property holds, i.e., an increase in $Q(\theta|\theta^{(t)})$ leads to an increase in $\ell(\theta)$ for all $\theta, \theta^{(t)} \in [a, b]$.

Proof: Consider the Taylor expansion of the $\ell(\theta)$ in a neighborhood of θ', we have

$$\ell(\theta)=\ell(\theta\prime)+(\theta-\theta\prime)^{\mathrm{T}}\nabla\ell(\theta\prime)+0.5(\theta-\theta\prime)^{\mathrm{T}}\nabla^2\ell(\theta^*)(\theta-\theta\prime)$$

for all $\theta, \theta' \in \mathbb{R}^q$ and some point θ^* between θ and θ'. Assumption (2.2) guarantees

$$\ell(\theta) - Q(\theta|\theta\prime)=0.5(\theta-\theta\prime)^{\mathrm{T}}[\nabla^2\ell(\theta^*)+B](\theta-\theta\prime) \geq 0, \quad \forall\theta,\theta\prime \in \mathbb{R}^q. \tag{2.7}$$

Thus, $\ell(\theta) - Q(\theta|\theta')$ achieves its minimum zero at $\theta = \theta'$. Since $[a, b]$ is a convex subset of $\mathbb{R}^q$, it follows that (2.7) holds for all $\theta, \theta' \in [a, b]$. Together with the condition $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})$ for all $\theta^{(t)}, \theta^{(t+1)} \in [a, b]$, we immediately obtain

$$\begin{aligned}\ell(\theta^{(t+1)}) &= \left\{\ell(\theta^{(t+1)}) - Q(\theta^{(t+1)}|\theta^{(t)})\right\} + Q(\theta^{(t+1)}|\theta^{(t)}) \\ &\geq \left\{\ell(\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})\right\} + Q(\theta^{(t)}|\theta^{(t)}) \\ &= \ell(\theta^{(t)}),\end{aligned}$$

where the inequality is strict if $\theta^{(t+1)} \neq \theta^{(t)}$.

2.3 The fastest QLB algorithm

As the EM algorithm (Meng and Rubin, 1991), the original QLB algorithm is also a linear iterative algorithm. Böhning and Lindsay (1988) showed that the original QLB algorithm converges linearly with matrix rate of convergence given by

$$\nabla M_B(\widehat{\theta})=I_q - B^{-1}[-\nabla^2\ell(\widehat{\theta})]=B^{-1}A, \tag{2.8}$$

where $A \equiv \nabla^2\ell(\hat{\theta}) + B \geq 0$. Similar to Meng (1994), we call the largest eigenvalue $\rho\{\nabla M_B(\hat{\theta})\}$ of $\nabla M_B(\hat{\theta})$ the global rate of convergence. In some literature, $\rho\{\nabla M_B(\hat{\theta})\}$ is also called the spectral radius of $\nabla M_B(\hat{\theta})$ (Fessler *et al.*, 1993). Böhning and Lindsay (1988) also showed that for the original QLB algorithm, the global rate of convergence $\rho\{\nabla M_B(\hat{\theta})\} \in [0, 1)$. Since the larger the value of $\rho\{\nabla M_B(\hat{\theta})\}$ the slower the algorithm, we usually call the smallest eigenvalue $s\{I - \nabla M_B(\hat{\theta})\} = 1 - \rho\{\nabla M_B(\hat{\theta})\}$ of $I - \nabla M_B(\hat{\theta})$ the global speed of the algorithm.

Suppose that there exist two GM matrices B_1 and B_2 which lead to two QLB algorithms with $\nabla M_{B_1}(\hat{\theta})$ and $\nabla M_{B_2}(\hat{\theta})$, respectively. Proposition 2 below tells that if B_1 majorizes B_2, then B_2 is preferred.

Proposition 2—*Suppose there exist two positive definite matrices B_1 and B_2 such that $B_1 \geq B_2 \geq -\nabla^2\ell(\theta)$ for all $\theta \in \mathbb{R}^q$. We have*

$$s\left\{I - \nabla M_{B_2}(\widehat{\theta})\right\} \geq s\{I - \nabla M_{B_1}(\widehat{\theta})\}.$$

That is, the QLB algorithm based on B_2 converges faster than the one based on B_1. Furthermore, if $B_1 > B_2 \geq -\nabla^2\ell(\theta)$ for all $\theta \in \mathbb{R}^q$, then $s\{I - \nabla M_{B_2}(\theta')\} > s\{I - \nabla M_{B_1}(\theta')\}$.

Proof: Our proof is similar to the proofs of Lemma 1 in Fessler *et al.* (1993) and Theorem 1 in Meng and van Dyk (1997). From (2.8), we have $I - \nabla M_{B_k}(\widehat{\theta})=B_k^{-1}H$, where $H \equiv -\nabla^2\ell(\hat{\theta})$ and $k = 1, 2$. Note that $B_1 \geq B_2$ implies $B_2^{-1} \geq B_1^{-1}$. If $H > 0$, then we immediately get

$$H^{1/2}B_2^{-1}H^{1/2} \geq H^{1/2}B_1^{-1}H^{1/2}.$$

Therefore, the smallest eigenvalue

$$s\left\{I-\nabla M_{B_2}(\widehat{\theta})\right\}=\inf_{z\neq 0}\left\{z^{\mathrm{T}}[I-\nabla M_{B_2}(\widehat{\theta})]z/z^{\mathrm{T}}z\right\}$$

satisfies:

$$\begin{aligned} s\left\{I-\nabla M_{B_2}(\widehat{\theta})\right\} &= s\left\{B_2^{-1}H\right\}=s\left\{H^{1/2}B_2^{-1}H^{1/2}\right\} \\ &\geq s\left\{H^{1/2}B_1^{-1}H^{1/2}\right\} \\ &= s\left\{B_1^{-1}H\right\} \\ &= s\left\{I-\nabla M_{B_1}(\widehat{\theta})\right\}. \end{aligned}$$

If $|H| = 0$, then $s\{I - \nabla M_{B_k}(\hat{\theta})\} = 0$. The proof for $B_1 > B_2 \geq H$ is similar.

We introduce a new notion before we derive the fastest QLB algorithm. A matrix B (> 0) is said to be the *smallest* global majorization matrix if (i) B is a GM matrix; and (ii) there does not exist a $B' > 0$ such that $B \geq B' \geq -\nabla^2\ell(\theta)$ for all $\theta \in \mathbb{R}^q$. First, the smallest GM matrix is unique (i.e., if there exists a $B'' > 0$ such that $B \geq B'' \geq -\nabla^2\ell(\theta)$ for all $\theta \in \mathbb{R}^q$, then $B'' = B$). Second, the smallest GM matrix is a GM, but the inverse is not true. Third, if B is the smallest GM matrix, then there exist infinite many GM matrices that majorize B (e.g., $r\,B \geq B$ for any $r \geq 1$). Proposition 3 below indicates that the fastest QLB algorithm corresponds to the one with the smallest GM matrix.

Proposition 3—*Let B be the smallest GM matrix and $r \geq 1$. The global speed of convergence for the QLB algorithm based on the GM matrix $r\,B$ is then given by*

$$s\left\{I-\nabla M_{rB}(\widehat{\theta})\right\}=r^{-1}\cdot s\left\{I-\nabla M_B(\widehat{\theta})\right\},$$

which is a monotonic decreasing function of r and its maximum is achieved at $r = 1$.

Proof: From (2.8), we have

$$\nabla M_{rB}(\widehat{\theta})=I_q-(rB)^{-1}[-\nabla^2\ell(\widehat{\theta})]=I-r^{-1}[I-\nabla M_B(\widehat{\theta})].$$

Since $\rho\{\nabla M_B(\hat{\theta})\} \in [0,1)$, we obtain

$$\rho\left\{\nabla M_{rB}(\widehat{\theta})\right\}=1-r^{-1}[1-\rho\left\{\nabla M_B(\widehat{\theta})\right\}]\in[1-r^{-1},1).$$

The result follows immediately from the definition of the global speed of convergence.

3. Extension to penalized problems

In this section, we generalize the original QLB algorithm to the penalized problem (1.6). That is,

$$\tilde{\theta}=\arg\max\,\ell_\lambda(\theta)=\arg\max\,\{\ell(\theta)-\lambda J_1(\theta)\}. \tag{3.1}$$

The penalized MLE $\hat{\theta}$ is the posterior mode (or maximum a posteriori estimate) in a Bayesian framework if we treat $c \cdot e^{-\lambda J(\theta)}$ as a prior density of θ. The QLB algorithm suggests that the $\tilde{\theta}$ can be obtained by iteratively calculating

$$\theta^{(t+1)}=\arg\max\,Q_\lambda(\theta|\theta^{(t)})=\arg\max\,\left\{Q(\theta|\theta^{(t)})-\lambda J_1(\theta)\right\}, \tag{3.2}$$

where Q is defined in (2.3). Similarly, we have the following results.

Proposition 4

Let the sequence $\{\theta^{(t)}\}_{t=0}^{\infty}$ *be generated by* (3.2). *Hence, we have (i)* $\ell_\lambda(\theta) - Q_\lambda(\theta|\theta^{(t)})$ *achieves its minimum (i.e., zero) at* $\theta = \theta^{(t)}$; *(ii) the ascent property holds, i.e., an increase in* $Q_\lambda(\theta|\theta^{(t)})$ *leads to an increase in* $\ell_\lambda(\theta)$ *for all* $\theta \in \mathbb{R}^q$.

Proof—Since $\ell_\lambda(\theta) - Q_\lambda(\theta|\theta^{(t)}) = \ell(\theta) - Q(\theta|\theta^{(t)})$, assertion (i) follows immediately. From (3.2), we have $Q_\lambda(\theta^{(t+1)})|\theta^{(t)}) \geq Q_\lambda(\theta^{(t)}|\theta^{(t)})$ for $\theta^{(t)}, \theta^{(t+1)} \in \mathbb{R}^q$. Thus,

$$\begin{aligned} \ell_\lambda(\theta^{(t+1)}) &= \left\{\ell_\lambda(\theta^{(t+1)}) - Q_\lambda(\theta^{(t+1)}\big|\theta^{(t)})\right\} + Q_\lambda(\theta^{(t+1)}\big|\theta^{(t)}) \\ &\geq \left\{\ell_\lambda(\theta^{(t)}) - Q_\lambda(\theta^{(t)}\big|\theta^{(t)})\right\} + Q_\lambda(\theta^{(t)}\big|\theta^{(t)}) \\ &= \ell_\lambda(\theta^{(t)}), \end{aligned}$$

where the inequality is strict if $\theta^{(t+1)} \neq \theta^{(t)}$.

Proposition 5

If $J_1(\theta)$ *is twice continuously differentiable and convex, then the QLB algorithm for the penalized problem in* (3.1) *converges with the matrix rate of convergence*

$$\nabla M(\tilde{\theta}) = [B + \lambda \cdot \nabla^2 J_1(\tilde{\theta})]^{-1} \tilde{A}, \qquad \text{where} \qquad \tilde{A} \equiv \nabla^2 \ell(\tilde{\theta}) + B \geq 0, \tag{3.3}$$

and the global rate of convergence $\rho\{\nabla M(\tilde{\theta})\} \leq \rho\{B^{-1}\tilde{A}\} < 1$.

Proof—Let the sequence $\{\theta^{(t)}\}$ be generated by the QLB algorithm in (3.2), which in fact defines a mapping $\theta \to M(\theta)$ from $\mathbb{R}^q$ to $\mathbb{R}^q$ such that $\theta^{(t+1)} = M(\theta^{(t)})$ for $t = 0, 1, \ldots, +\infty$. From (3.2), $\theta^{(t+1)}$ can be calculated by differentiating $Q(\theta|\theta^{(t)}) - \lambda J_1(\theta)$ with respect to θ and then setting to zero, i.e.,

$$\nabla \ell(\theta^{(t)}) - B(\theta - \theta^{(t)}) - \lambda \nabla J_1(\theta) = 0. \tag{3.4}$$

Noting that $\theta = M(\theta^{(t)})$, we differentiate (3.4) with respect to $\theta^{(t)}$ instead, and obtain

$$\nabla^2 \ell(\theta^{(t)}) - B(\nabla M(\theta^{(t)}) - I_q) - \lambda \nabla^2 J_1(\theta) \cdot \nabla M(\theta^{(t)}) = 0.$$

At convergence (i.e., $\theta = \theta^{(t)} = \tilde{\theta}$), we have $\nabla M(\tilde{\theta}) = [B + \lambda \cdot \nabla^2 J_1(\tilde{\theta})]^{-1}(\nabla^2 \ell(\tilde{\theta}) + B)$, and (3.3) follows. Since $\nabla^2 J(\tilde{\theta}) \geq 0$, by Proposition (a) in Green (1990), we have $\rho\{\nabla M(\tilde{\theta})\} \leq \rho\{B^{-1}\tilde{A}\} < 1$.

It is noteworthy that when $\lambda = 0$, we see that (3.3) reduces to (2.8). Hence, Proposition 5 implies that the QLB algorithm for the penalized problem converges faster than the QLB algorithm for the unpenalized problem.

When $J_1(\theta)$ is not totally differentiable (see, the lasso regression (6.1) discussed later), the penalized problem in (3.1) can sometimes be reformulated as

$$\tilde{\theta} = \arg \max_{\theta \in [a,b]} \{\ell(\theta) - \lambda J_2(\theta)\}, \tag{3.5}$$

where $J_2(\theta)$ is differentiable everywhere. Thus, the QLB algorithm suggests that the $\tilde{\theta}$ in (3.5) can be obtained by the following iteration

$$\theta^{(t+1)} = \arg \max_{\theta \in [a,b]} \left\{Q(\theta\big|\theta^{(t)}) - \lambda J_2(\theta)\right\}. \tag{3.6}$$

4. A pseudo-Newton method

Like EM-type algorithms, the QLB algorithm is often criticized for its slow convergence in some applications, especially for solving complicated problems or in high-dimensional data

analysis. In this section, we develop a pseudo-Newton method that on the one hand maintains the simplicity of the proposed QLB algorithm and on the other hand achieves the speed of convergence of the Newton method.

The Newton method solves (2.1) using the following iteration

$$\theta^{(t+1)}=\arg \min_{\theta\in[a,b]} \|\xi(-\nabla^2\ell(\theta^{(t)}),\theta^{(t)}) - C^{(t)}\theta\|^2, \tag{4.1}$$

where $\xi(\cdot, \theta^{(t)})$ is defined by (2.5), and the Cholesky decomposition $-\nabla^2\ell(\theta^{(t)}) = C^{(t)\mathrm{T}} C^{(t)}$ has to be calculated at each iteration. The scoring method simply replaces the observed information by the expected information. If we replace $-\nabla^2\ell(\theta^{(t)})$ in (4.1) with a *surrogate matrix* $\widehat{B}^{\mathrm{U}}>0$, then a pseudo-Newton algorithm can be defined by the following iteration

$$\theta^{(t+1)}=\arg \min_{\theta\in[a,b]} \|\xi(\widehat{B}^{\mathrm{U}},\theta^{(t)}) - C_*\theta\|^2, \tag{4.2}$$

where $\widehat{B}^{\mathrm{U}}=C_*^{\mathrm{T}}C_*$ is required to be calculated only once. Let $\widehat{\theta}^{\mathrm{U}}$ denote the unconstrained MLE of θ in (2.1). If $-\nabla^2\ell(\widehat{\theta}^{\mathrm{U}})>0$, set $\widehat{B}^{\mathrm{U}}= -\nabla^2\ell(\widehat{\theta}^{\mathrm{U}})$; otherwise a minor modification of $-\nabla^2\ell(\widehat{\theta}^{\mathrm{U}})$ can be served as $\widehat{B}^{\mathrm{U}}$. For logistic models, one may refer to (5.3). The following proposition shows that the pseudo-Newton (4.2) converges faster than the fastest QLB algorithm.

Proposition 6

Let B be the smallest GM matrix. (i) If $B \geq \widehat{B}^{\mathrm{U}}>0$,then $s\left\{I - \nabla M_{\widehat{B}^{\mathrm{U}}}(\widehat{\theta})\right\} \geq s\left\{I - \nabla M_B(\widehat{\theta})\right\}$, *i.e., the pseudo-Newton algorithm converges faster than the fastest QLB algorithm based on B; and (ii) If* $B>\widehat{B}^{\mathrm{U}}>0$,then $s\left\{I - \nabla M_{\widehat{B}^{\mathrm{U}}}(\widehat{\theta})\right\}>s\left\{I - \nabla M_B(\widehat{\theta})\right\}$.

Proof—Using the condition $B_2 \geq -\nabla^2\ell(\theta)$ in the proof of Proposition 2, Proposition 6 follows immediately.

5. Application to logistic regression with constraints

5.1 The fastest QLB and the pseudo-Newton algorithms

The key to the QLB algorithm is to find a GM matrix satisfying condition (2.2). For the logistic model (1.3), Böhning and Lindsay (1988) gave the smallest GM matrix. From (1.4), the score and the observed information are given by

$$\begin{aligned} \nabla\ell(\theta) &= \textstyle\sum_{i=1}^m (y_i - n_i p_i)x_{(i)} = X^{\mathrm{T}}(y - N_p), \qquad \text{and} \\ -\nabla^2\ell(\theta) &= \textstyle\sum_{i=1}^m n_i p_i(1 - p_i)x_{(i)}x_{(i)}^{\mathrm{T}} = X^{\mathrm{T}}NDX, \end{aligned}$$

respectively, where $X^{\mathrm{T}} = (x_{(1)},\ldots,x_{(m)})$, $y = (y_1,\ldots,y_m)^{\mathrm{T}}$, $p = (p_1,\ldots,p_m)^{\mathrm{T}}$, $N=\mathrm{diag}(n_1,\ldots,n_m)$, and $D=\mathrm{diag}(p_1(1 - p_1),\ldots,p_m(1 - p_m))$. (5.1)

For each i, since $0.25 \geq p_i(1 - p_i)$, the smallest GM matrix is

$$B=(1/4)X^{\mathrm{T}}NX, \tag{5.2}$$

which corresponds to the fastest QLB algorithm. On the other hand, let $\widehat{\theta}^{\mathrm{U}}$ denote the unconstrained MLE of θ in (1.4),

$$\widehat{p}_i^{\mathrm{U}}=\exp\left\{x_{(i)}^{\mathrm{T}}\widehat{\theta}^{\mathrm{U}}\right\}/[\,1+\exp\left\{x_{(i)}^{\mathrm{T}}\widehat{\theta}^{\mathrm{U}}\right\}],$$

and $\widehat{d}_i^{\mathrm{U}}=\widehat{p}_i^{\mathrm{U}}(1-\widehat{p}_i^{\mathrm{U}})$ if $\widehat{p}_i^{\mathrm{U}}\in(0,1)$ and $\widehat{d}_i^{\mathrm{U}}=0.25$ otherwise ($i=1,\ldots,m$). Since the conditions in Proposition 6 are satisfied (i.e., $B>\widehat{B}^{\mathrm{U}}>0$), setting

$$\widehat{B}^{\mathrm{U}}=X^{\mathrm{T}}N\,\mathrm{diag}(\widehat{d}_i^{\mathrm{U}},\ldots,\widehat{d}_m^{\mathrm{U}})X \tag{5.3}$$

yields the pseudo-Newton algorithm in (4.2).

5.2 Standard errors

Using the aforementioned efficient algorithms for calculating $\hat{\theta}$, calculating the standard errors of $\hat{\theta}$ via bootstrapping becomes computationally feasible (Efron and Tibshirani, 1993). Having obtained the restricted MLE $\hat{\theta}$ from (1.5) by the fastest QLB algorithm or the pseudo-Newton algorithm, we can directly generate a bootstrap sample $\{y_i^*\}_{i=1}^m$ with $y_i^*\overset{\text{ind}}{\sim}$ Binomial $(n_i,\ \exp\{x_{(i)}^{\mathrm{T}}\widehat{\theta}\}/[1+\exp\{x_{(i)}^{\mathrm{T}}\widehat{\theta}\}])$ and compute the corresponding bootstrap replication $\hat{\theta}^*$. Independently repeating this process G times, we obtain G bootstrap replications $\{\widehat{\theta}^*(g)\}_{g=1}^G$ with $\widehat{\theta}^*(g)=(\widehat{\theta}_1^*(g),\ldots,\widehat{\theta}_q^*(g))^{\mathrm{T}}$. Therefore, the standard error $\mathrm{se}(\widehat{\theta}_j)$ of $\widehat{\theta}_j$ can be estimated by the sample standard deviation of the G replications.

5.3 Simulations and comparisons: Binomial model with simplex constraints

Liu (2000) used the EM algorithm to find the MLE of $p=(p_1,\ldots,p_m)^{\mathrm{T}}$ in the binomial model (1.3) with simplex constraints $p_i=\sum_{j=1}^q f_{ij}\alpha_j$, where $\{f_{ij}\}$ are known and nonnegative, $\alpha_j\geq 0$, $1\leq i\leq m$, $1\leq j\leq q$, and $\sum_{j=1}^q\{\max_{1\leq i\leq m}(f_{ij})\cdot\alpha_j\}\leq 1$. Noticing the non-negativity assumption on the entries $\{f_{ij}\}$, we can immediately conclude that his approach is inapplicable to, for instance, umbrella, tree, increasing concave, sigmoid and bell-shaped orderings (see, Robertson, *et al.*, 1988; Schmoyer, 1984; Meyer, 1999) since some entries in the transformation matrices for these orderings are negative.

Here, we can solve these problems by considering an equivalent optimization problem via the transformation $\mu_i=\mathrm{logit}(p_i)=\log[p_i/(1-p_i)]$. For example, for problems with umbrella region $S(p)=\{p: p_1\leq\ldots\leq p_h\geq p_{h+1}\geq\ldots\geq p_m,\ 0\leq p_i\leq 1\}$, finding

$$\widehat{p}=\arg\max_{p\in S(p)}\sum_{i=1}^m\{y_i\log p_i+(n_i-y_i)\log(1-p_i)\} \tag{5.4}$$

is equivalent to finding $\widehat{\mu}=\arg\max_{\mu\in S(\mu)}\sum_{i=1}^m\{y_i\mu_i-n_i\log(1+e^{\mu_i})\}$, where $S(\mu)=\{\mu:\mu_1\leq\ldots\leq\mu_h\geq\mu_{h+1}\geq\ldots\geq\mu_m,\ \mu_i\in\mathbb{R}^1\}$. Let $\mu=X\theta$. It is easy to see that finding $\hat{\mu}$ is equivalent to solving (1.5) with $\theta_{m\times 1}\in[a,b]=\mathbb{R}\times\mathbb{R}_+^{m-1}$ and

$$X_{m\times m}=\begin{pmatrix}\Delta_h & O\\ 1_{m-h}1_h^{\mathrm{T}} & -\Delta_{m-h}\end{pmatrix}, \tag{5.5}$$

where $\Delta_h=(\delta_{ij})$ is a $h\times h$ matrix with $\delta_{ij}=1$ for $i\geq j$ and $\delta_{ij}=0$ for $i<j$. If $\hat{\theta}$ can be obtained from (1.5), we have $\widehat{p}_i=\exp(x_{(i)}^{\mathrm{T}}\widehat{\theta})/\{1+\exp(x_{(i)}^{\mathrm{T}}\widehat{\theta})\}$.

We compare the fastest QLB algorithm based on (5.2) and the pseudo-Newton algorithm based on (5.3) with the existing algorithm via a simulated data set. Let $m=40$ and $\{(n_i,p_i)\}_{i=1}^m$ be given in Table 1, where $\{p_i\}_{i=1}^m$ are assumed to be restricted by the umbrella ordering with $h=25$. From (1.3), we generate independent binomial samples $\{y_i\}_{i=1}^m$ and report them in Table 1. The objective is to find $\hat{p}$ in (5.4). Starting with $\theta^{(0)}=(1,\ldots,1)^{\mathrm{T}}$, the two algorithms converged to the maximum point $\hat{p}$ which is reported in the 5-th and the 11-th column of Table 1 with the

log-likelihood showing a steady increase to its maximum value of −69.2 (see Figure 1(a)). The fastest QLB took 60 iterations with a CPU time of 9.969 seconds for the log-likelihood to achieve this final value, while the pseudo-Newton algorithm took 25 iterations with a CPU time of 3.766 seconds. The corresponding standard errors are calculated by the parametric bootstrapping with 1,000 replications via the fastest QLB algorithm (see Table 1). Figure 1(b) and Figure 1(c) compare the true proportions p_i (the 3-rd and the 9-th column in Table 1), y_i/n_i and the constrained estimates $\widehat{p}_i$.

To check the correctness of the fastest QLB algorithm, we consider the algorithm by Dykstra (1983). Theorem 3.1 of Barlow and Brunk (1972) showed that finding (5.4) is equivalent to finding the weighted LSE

$$\widehat{p}^{\mathrm{WLSE}}=\arg \quad \min \sum_{i=1}^{m} n_i(y_i/n_i - p_i)^2 \tag{5.6}$$

subject to the same constraints $p \in S(p)$. Let $p = X\theta$, where $\theta \in [0,1]^m$ and X is given by (5.5). Hence, $\widehat{p}^{\mathrm{WLSE}}=X\widehat{\theta}^{\mathrm{WLSE}}$ and

$$\widehat{\theta}^{\mathrm{WLSE}}=\arg \min_{\theta\in[0,1]^m} \|y^* - X^*\theta\|^2, \tag{5.7}$$

where $y^*=N^{1/2}(\frac{y_1}{n_1},\ldots,\frac{y_m}{n_m})^{\mathrm{T}}$, $X^* = N^{1/2}X$ and N is given by (5.1). Applying Dykstra algorithm to (5.7), we obtained $\widehat{p}^{\mathrm{WLSE}}$ with a CPU time of 4.65 seconds. Figure 1(d) shows that the solutions from the fastest QLB and Dykstra algorithm are identical.

5.4 Example 1: Down syndrome data

The incidence of *Down syndrome* (DS) is highly dependent on maternal age. A large scale study for ascertainment of DS cases in Massachusetts from 1958 to 1965 was conducted and the data are given in Table 2 (Hook and Fabia, 1978).

Let n_i, y_i and p_i denote the number of live births, the number of DS cases and the probability of DS (i.e., the DS incidence) within the maternal age class i ($i = 1, \ldots, m$ with $m = 35$). Let z_i represent the average maternal age within the age class i ("$i = 1$" is corresponding to average age 15.562 and "$i = 35$" corresponding to average age 49.41). Define $\mu_i = \mathrm{logit}\,(p_i) = f(z_i)$, $i = 1, \ldots, m$. Figure 2(d) plots $\mathrm{logit}\,(\frac{y_i}{n_i})$ marked with "•" against z_i (except for those cases with zero frequency) and it suggests that $f(\cdot)$ may be a non-decreasing and convex curve. Following the suggestion in Geyer (1991), we restrict μ by

$$0 \le \frac{\mu_2 - \mu_1}{z_2 - z_1} \le \frac{\mu_3 - \mu_2}{z_3 - z_2} \le \cdots \le \frac{\mu_m - \mu_{m-1}}{z_m - z_{m-1}},$$

and obtain $\mu = X\theta$, where

$$X_{m\times m}=\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & z_2 - z_1 & 0 & \cdots & 0 & 0 \\ 1 & z_3 - z_1 & z_3 - z_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & z_{m-1} - z_1 & z_{m-1} - z_2 & \cdots & z_{m-1} - z_{m-2} & 0 \\ 1 & z_m - z_1 & z_m - z_2 & \cdots & z_m - z_{m-2} & z_m - z_{m-1} \end{pmatrix} \tag{5.8}$$

and $\theta_{m\times 1} \in \mathbb{R} \times \mathbb{R}_+^{m-1}$. Therefore, finding the constrained $\hat{p}$ is equivalent to finding $\hat{\theta}$ in (1.5) with $[a,b]=\mathbb{R} \times \mathbb{R}_+^{m-1}$ and X being given by (5.8).

Starting with $\theta^{(0)} = (-1, 0.1,\ldots,0.1)^{\mathrm{T}}$, the fastest QLB took 1, 500 iterations to reach the log-likelihood value of −104.115 with the CPU time of 262.938 seconds. On the contrary, the

pseudo-Newton method converged to the maximum log-likelihood value −104.069 in 14 iterations with 3.703 seconds. Table 3 reports the intermediate results of the pseudo-Newton algorithm for this problem. This example numerically justifies Proposition 6 and demonstrates that the pseudo-Newton algorithm can be 71 or 107 times faster than the fastest QLB algorithm in terms of CPU time or numbers of iteration at the expense of lacking the warrant of automatic monotone convergence.

The $\hat{p}$ are obtained by $\widehat{p}_i = e^{x_{(i)}^{\mathrm{T}}\widehat{\theta}}/(1+e^{x_{(i)}^{\mathrm{T}}\widehat{\theta}})$. The standard errors are calculated by the parametric bootstrap with 1000 replications via the pseudo-Newton algorithm (see Table 2). Figure 2(a)–(c) show the comparisons of log-likelihoods for the two algorithms. Two fitted curves obtained via the two algorithms are showed in Figure 2(d).

6. Application to lasso logistic regression

We apply the fastest QLB algorithm to select significant variables in logistic regression models. Note that the lasso solution (1.1) is a special case of (1.6) or (3.1) with non-differentiable penalty at the origin. Since the lasso estimates and the unconstrained MLEs share signs, lasso regression is equivalent to quadratic optimization with non-negative constraints, thus avoiding the problem of non-differentiable penalty function. Therefore, the M-step of the fastest QLB algorithm can be readily implemented using the built-in S-PLUS function nnls.fit (nonnegative least squares).

6.1 The L_1-penalty

Let $\ell(\theta)$ be given by (1.4). Obviously, finding (1.1) is equivalent to finding

$$\widehat{\theta}^{\text{lasso}} = \arg\max \left\{ \ell(\theta) - \lambda \cdot \sum_{j=1}^{q} |\theta_j| \right\}, \tag{6.1}$$

where $\lambda > 0$ is a smoothing parameter. Note that (6.1) is a special case of (3.1). The fastest QLB algorithm can be applied to obtain $\widehat{\theta}^{\text{lasso}}$ by iteratively computing

$$\theta^{(t+1)} = \arg \min \left\{ \|\xi(B,\theta^{(t)}) - C\theta\|^2 + \lambda \cdot \sum_{j=1}^{q} |\theta_j| \right\}, \tag{6.2}$$

where $\xi(B, \theta^{(t)})$ and B are defined in (2.5) and (5.2), respectively.

Let $\widehat{\theta}^{\text{U}}$ denote the unconstrained MLE of θ in the logistic model (1.4) and $\upsilon = (\upsilon_1, \ldots, \upsilon_q)^{\mathrm{T}}$ be its sign vector (i.e., $\upsilon_j = \text{sign}(\widehat{\theta}_j^{\text{U}}) = +1, 0, \text{or } -1$ corresponding to positive, zero, or negative values of $\widehat{\theta}_j^{\text{U}}$). The geometric property of the lasso solution suggests that both $\widehat{\theta}^{\text{lasso}}$ and $\widehat{\theta}^{\text{U}}$ share signs (see, Efron *et al.* 2004, Lemma 7 and 8). This implies the lasso estimator $\widehat{\theta}^{\text{lasso}} \in \left\{ (\upsilon_1\beta_1, \ldots, \upsilon_q\beta_q)^{\mathrm{T}} = \text{diag}(\upsilon)\beta : \beta \in \mathbb{R}_+^q \right\}$ Given $\theta^{(t)}$ and from (6.2), we have

$$\theta^{(t+1)} = \text{diag}(\upsilon) \cdot \beta^{(t+1)}, \quad \text{and} \tag{6.3}$$

$$\beta^{(t+1)} = \arg \min_{\beta \in \mathbb{R}_+^q} \|\eta(B,\theta^{(t)}) - Z\beta\|^2, \tag{6.4}$$

where $\eta(B, \theta^{(t)}) = (Z^{\mathrm{T}})^{-1}[\text{diag}(\upsilon C^{\mathrm{T}}\xi(B, \theta^{(t)}) - 0.5\lambda \cdot 1]$ and Z can be obtained via the Cholesky decomposition such that $\text{diag}(\upsilon)C^{\mathrm{T}}C\text{diag}(\upsilon) = Z^{\mathrm{T}}Z$. Since (6.4) is a quadratic optimization problem with non-negative constraints, we can utilize the built-in S-PLUS function nnls.fit to solve (6.4) iteratively.

6.2 Automatic choice of the smoothing parameter via GCV

The optimal smoothing parameter $\widehat{\lambda}^{\text{opt}}$ can be selected automatically via minimizing an approximate *generalized cross-validation* (GCV) statistic (Craven and Wahba, 1979). For any given $\lambda > 0$, we calculate the lasso estimate based on (6.3) and (6.4) and denote it by $\widehat{\theta}_{\lambda}^{\text{lasso}}$. The GCV statistic is defined as

$$\text{GCV}(\lambda) = -\frac{\ell(\widehat{\theta}_{\lambda}^{\text{lasso}})}{m[1 - e(\lambda)/m]^2},$$

where

$$e(\lambda) = \text{tr}[X(X^{\text{T}}NDX + \lambda W^{-})^{-1}X^{\text{T}}ND]$$

is the effective number of parameters, W^{-} denotes the Moore-Penrose generalized inverse of $W = \text{diag}(|\widehat{\theta}_{\lambda}^{\text{lasso}}|)$, N and D are defined in (5.1). We determine the optimal $\widehat{\lambda}^{\text{opt}}$ by minimizing GCV(λ) over a grid of λ values.

6.3 Example 2: Kyphosis data

This data set consists of retrospective measurements on 83 laminectomy patients (Hastie and Tibshirani, 1990, p.282). The outcome is the status of kyphosis (1 = present, 0=absent). The predictors include: x_1 = age in months at time of the operation, x_2 = number of vertebrae levels, and x_3 = starting vertebrae level. The goal is to identify risk factors for kyphosis. To explore possible non-linear effects of the risk factors, we include three main effects and three quadratic effects in the full model. To compare the fastest QLB algorithm with Tibshirani (1996) algorithm, we do not include the interaction effects. Since all the covariates are continuous, they are standardized individually in our analysis. The full logistic regression model is

$$\text{logit}\ \{\Pr(Y=1)\} = \theta_0 + \sum_{j=1}^{3}\theta_j x_j + \sum_{j=1}^{3}\theta_{3+j}x_j^2.$$

The SAS proc logistic with backward stepwise selection removed the x_2^2-term and yielded the following model

$$-2.6451 + 0.8310x_1 + 0.7955x_2 - 2.2670x_3 - 1.5320x_1^2 - 1.1533x_3^2. \quad (6.5)$$

To apply the fastest QLB algorithm in (6.3) and (6.4) to obtain the lasso solution $\widehat{\theta}^{\text{lasso}}$, we first need to calculate the unconstrained MLE. We obtain $\widehat{\theta}^{\text{U}} = (-2.6422, 0.8270, 0.7673, -2.2688, -1.5406, 0.0321, -1.1582)^{\text{T}}$ and its sign vector $\upsilon = (-1, 1, 1, -1, -1, 1, -1)^{\text{T}}$. The GCV method is used to select the optimal smoothing parameter. Figure 3(a) depicts the plot of GCV versus λ. The optimal $\widehat{\lambda}^{\text{opt}}$ is given by 0.351. Using $\theta^{(0)} = \upsilon$ as the initial values, the fastest QLB algorithm in (6.3) and (6.4) converged in $t = 80$ iteration with the CPU time of 1.687 seconds. Figure 3(b) shows the monotone convergence of the algorithm. The resulting lasso solution is

$$\widehat{\theta}^{\text{lasso}} = (-2.2640, 0.6581, 0.6919, -1.8373, -1.2517, 0.0000, -0.8589)^{\text{T}}, \quad (6.6)$$

which considerably coincides with (6.5). The corresponding standard errors with 1,000 bootstrap replications are 0.5208, 0.4419, 0.3984, 0.6636, 0.6622, (−), and 0.5095, respectively. However, the lasso estimates given by Tibshirani (1996) are $-1.42 + 0.03x_1 + 0.31x_2 - 0.48x_3 - 0.28x_1^2$, which are different from (6.5) and (6.6). To some extent, this is expected. First, the lasso algorithm in his paper was based on the inequality constraint in (1.1), while our QLB algorithm is based on the penalized optimization with non-negativity constraints. In addition, Tibshirani showed that different criteria (e.g., CV, GCV

and Stein unbiased estimate of risk) could result in different choices of the tuning parameter u. It is not clear which one was used in his paper.

6.4 Example 3: Colon microarray data

The data set is composed of 2000 genes in 22 normal colon tissue samples and 40 tumor colon samples (Alon *et al.*, 1999). The outcome is binary (1 = tumor colon, 0 = normal colon). For $s = 1, \ldots, 2000$, we fit marginal logistic models with the expression levels for the s-th gene as a one-dimensional covariate. All genes with marginal p-values less than 0.001 are included in the second step logistic model fittinig. Twenty five out of 2000 genes are identified to be marginally significant at the 0.001 level. Since the sample size $m = 62$ is larger than $q = 25$, the number of variables, we first calculate the unconstrained MLE $\widehat{\theta}^{\mathrm{U}}$ which is given by

$$\begin{array}{rrrrrrrrr} (-9.908, & 3.795, & 18.414, & 5.191, & -3.800, & 13.509, & -33.823, & -3.371, & 0.182, \\ 6.184, & -9.619, & 7.296, & 9.060, & 4.765, & -27.769, & 4.486, & -3.973, & 10.797, \\ -4.303, & 16.403, & 5.354, & 5.136, & 1.811, & -2.696, & -1.708, & -17.991)^{\mathrm{T}}. & \end{array}$$

Thus, we can obtain the corresponding sign vector $\upsilon = (-1, 1, \ldots, -1)^{\mathrm{T}}$. The GCV criterion is used to select the optimal smoothing parameter. Figure 4 depicts the plot of GCV versus λ. The optimal $\widehat{\lambda}^{\mathrm{opt}}$ is 0.0691 and the corresponding GCV is 0.1745. Using $\theta^{(0)} = \upsilon$ as the initial values, the fastest QLB algorithm in (6.3) and (6.4) converged to the following lasso solution

$$\begin{array}{rrrrrrrrr} (-1.983, & 0.832, & 2.345, & 0.347, & 0.000, & 1.996, & -2.063, & -2.695, & 0.000, \\ 0.548, & -1.808, & 1.595, & 0.987, & 0.314, & -2.922, & 0.919, & -0.766, & 0.718, \\ -1.009, & 2.019, & 0.999, & 0.538, & 0.871, & -0.638, & -0.489, & -1.525)^{\mathrm{T}}. & \end{array}$$

That is, 23 out of the 25 genes are identified under the GCV criterion.

7. Discussion

We developed an EM-type algorithm – the QLB algorithm – for estimating bounded parameters and selecting variables via L_1-penalty in logistic regression models. The key to the application of the QLB algorithm is to find a positive definite matrix that globally majorizes the observed information. To our knowledge, besides the logistic model, such a matrix (i.e., the smallest global majorization matrix) exists for both the Cox's model (Böhning and Lindsay, 1988) and the multinomial logistic model (Böhning, 1992; Kim *et al.*, 2006). Thus, the QLB algorithm can be used for constrained parameter estimation and variable selection for these models in which EM algorithm is not applicable because of lacking a missing-data structure. We showed that the smallest global majorization matrix corresponds to the fastest QLB algorithm. Furthermore, we proposed a pseudo-Newton algorithm that maintains both the simplicity of the QLB algorithm and the fast convergence of the Newton method. Our numerical examples in §5.3 and §5.4 showed that the pseudo-Newton algorithm is dramatically faster than the fastest QLB algorithm (up to 71 in CPU time or 107 times in numbers of iteration). It is worthwhile to investigate the existence of such a global majorization matrix in other models (e.g., binomial models with the complementary log-log link). However, when the dimension q is very large, the L-BFGS-B algorithm (a limited-memory algorithm for solving large nonlinear optimization problems subject to box constraints) of Zhu *et al.* (1997) may be one alternative.

It is well known that it is often not possible to make Bayesian inference analytically for logistic regression (Gilks, 1996) and the Gibbs sampling in conditional logistic regression does not converge (Mehta *et al.*, 2000, p.106–107). In contrast, the fastest QLB algorithm can be used to obtain the posterior mode with closed-form expression at each iteration. In fact, let the log-likelihood $\ell(\theta)$ be given by (1.4) and $N_q(0, V)$ be the prior density of θ. From (3.2), the posterior

mode can be obtained iteratively by calculating $\theta^{(t+1)} = \arg\max \{Q.(\theta|\theta^{(t)} - 0.5\theta^{T}V^{-1}\theta\} = (B + V^{-1})^{-1}[\Delta\ell(\theta^{(t)}) + B\theta^{(t)}]$.

In this paper, we only discuss the situation of independent binary data. It is worth-while to consider the QLB algorithm for logistic models with correlated or clustered binary data. Furthermore, for some generalized linear models (e.g., Poisson regression for counting data), a global majorization matrix does not exist and the QLB algorithm is not applicable. Therefore, it would be of interest to derive a similar fast algorithm to estimate constrained parameters and selecting variables in these generalized linear models. The SPLUS/R programs are available upon request from the authors.

Acknowledgments

GL Tian and M Tan's research was supported in part by U.S. National Cancer Institute grants CA119758 and CA106767. The work of ML Tang was fully supported by Hong Kong Baptist University grant FRG/06-07/II-20. The research of HB Fang was partially supported by U.S. National Cancer Institute grant CA106767.

References

Agresti, A. Categorical Data Analysis. Second Edition. New York: John Wiley & Sons; 2002.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 1999;96:6745–6750. [PubMed: 10359783]

Barlow RE, Brunk HD. The isotonic regression problem and its dual. Journal of the American Statistical Association 1972;67:140–147.

Böhning D. Multinomial logistic regression algorithm. Annals of the Institute of Statistical Mathematics 1992;44:197–200.

Böohning D, Lindsay BG. Monotonicity of quadratic approximation algorithms. Annals of the Institute of Statistical Mathematics 1988;40:641–663.

Collett, D. Modeling Binary Data. London: Chapman & Hall; 1991.

Cox, DR.; Oakes, D. Analysis of Survival Data. London: Chapman & Hall; 1984.

Craven P, Wahba G. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik 1979;31:377–403.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B 1977;39:1–38.

Dykstra RL. An algorithm for restricted least squares regression. Journal of the American Statistical Association 1983;78:837–842.

Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. Boca Raton: Chapman & Hall/CRC; 1993.

Efron B, Hastie T, Johnstone I, Tibshirani RJ. Least angle regression (with discussion). The Annals of Statistics 2004;32:407–499.

Fessler JA, Clinthorne NH, Rogers WL. On complete-data spaces for PET reconstruction algorithms. IEEE Transactions on Nuclear Science 1993;40(4):1055–1061.

Frank IE, Friedman JH. A statistical view of some chemometrics regression tools (with discussion). Technometrics 1993;35(2):109–148.

Geyer CJ. Constrained maximum likelihood exemplified by isotonic convex logistic regression. Journal of the American Statistical Association 1991;86:717–724.

Gilks, WR. Full conditional distributions. In: Gilks, WR.; Richardson, S.; Spiegelhalter, DJ., editors. Markov Chain Monte Carlo in Practice. London: Chapman & Hall; 1996. p. 75-88.

Green PJ. On the use of the EM algorithm for penalized likelihood estimation. Journal of the Royal Statistical Society, Series B 1990;52:443–452.

Hastie, TJ.; Tibshirani, RJ. Generalized Additive Models. Boca Raton: Chapman & Hall/CRC; 1990.

Hook EB, Fabia JJ. Frequency of Down syndrome in livebirths by single-year maternal age interval: Results of a Massachusetts study. Teratology 1978;17:223–228. [PubMed: 150062]

Khuri AI. A constrained least squares problem. Communications in Statistics: Simulation and Computation 1976;5:82–84.

Kim Y, Kwon S, Song SH. Multiclass spare logistic regression for classification of multiple cancer types using gene expression data. Computational Statistics and Data Analysis 2006;51:1643–1655.

Liu CH. Estimation of discrete distributions with a class of simplex constraints. Journal of the American Statistical Association 2000;95:109–120.

Mehta CR, Patel NR, Senchaudhuri P. Efficient Monte Carlo methods for conditional logistic regression. Journal of the American Statistical Association 2000;95:99–108.

Meng XL. On the rate of convergence of the ECM algorithm. The Annals of Statistics 1994;22:326–339.

Meng XL, Rubin DR. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. Journal of the American Statistical Association 1991;86:899–909.

Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 1993;80:267–278.

Meng XL, van Dyk D. The EM algorithm — an old folk-song sung to a fast new tune (with discussion). Journal of the Royal Statistical Society, Series B 1997;59:511–567.

Meyer MC. An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimensions. Journal of Statistical Planning and Inference 1999;81:13–31.

Robertson, T.; Wright, FT.; Dykstra, RL. Order Restricted Statistical Inference. New York: John Wiley & Sons; 1988.

Schmoyer RL. Sigmoidally constrained maximum likelihood estimation in quantal bioassay. Journal of the American Statistical Association 1984;79:448–453.

Silvapulle, MJ.; Sen, PK. Constrained Statistical Inference: Inequality, Order, and Shape Restrictions. New York: John Wiley & Sons; 2005.

Tan, M.; Tian, GL.; Fang, HB. Estimating restricted normal means using the EM-type algorithms and IBF sampling. In: Huang, J.; Zhang, H., editors. Development of Modern Statistics and Related Topics — In Celebration of Prof. Yaoting Zhang's 70th Birthday. New Jersey: World Scientific; 2003. p. 53-73.

Tan M, Tian GL, Fang HB, Ng KW. A fast EM algorithm for quadratic optimization subject to convex constraints. Statistica Sinica 2007;17(3):945–964.

Tibshirani RJ. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 1996;58:267–288.

Zhu CY, Byrd RH, Lu PH, Nocedal J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software 1997;23(4):550–560.

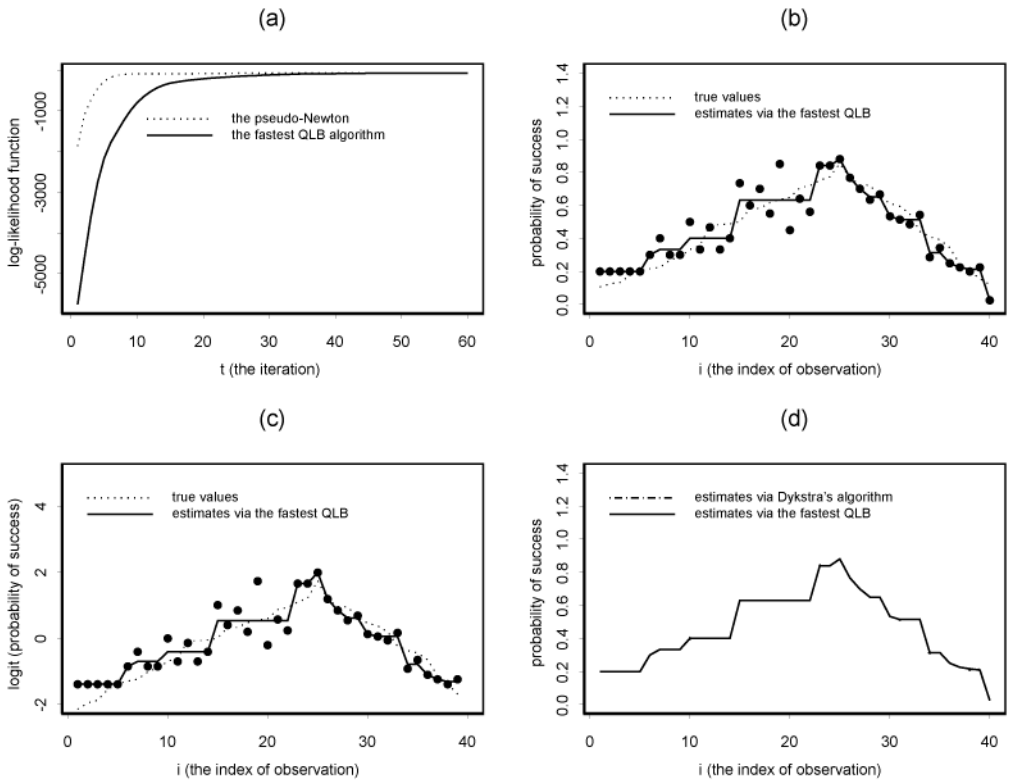

Figure 1.

Simulated data in Table 1. (a) Plots of the log-likelihood function against the iteration t for the pseudo-Newton algorithm based on (5.3) and the fastest QLB algorithm based on (5.2). (b) Comparison among the true proportions p_i (denoted by "…"), the unconstrained estimates y_i/n_i (denoted by "•"), and the constrained estimates $\widehat{p}_i$ (denoted by "—") obtained via the fastest QLB. (c) The same comparison as in (b) but with logit scale. (d) Comparison of the weighted LSE $\widehat{p}_i^{\mathrm{WLSE}}$ (denoted by "·–") in (5.6) via Dykstra algorithm with the constrained MLE $\widehat{p}_i$ (denoted by "—") in (5.4) via the fastest QLB.

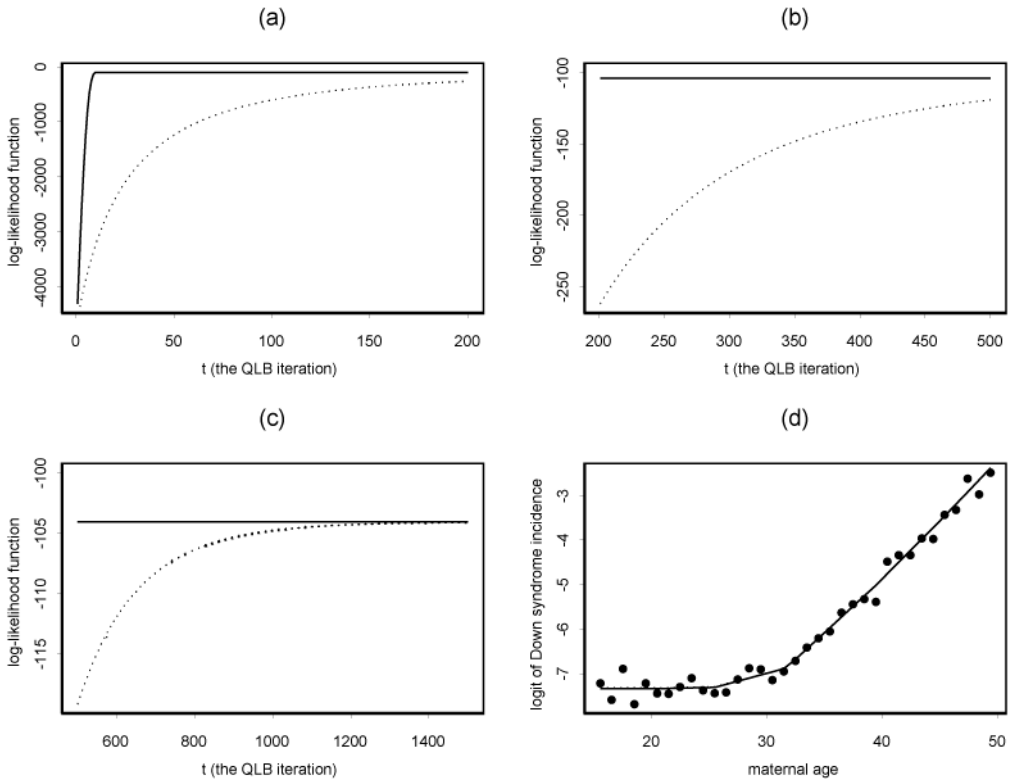

Figure 2.

The Down syndrome data in Table 2. (a)–(c) Comparison of the log-likelihoods for the fastest QLB algorithm based on (5.2) (denoted by "…") and the pseudo-Newton algorithm based on (5.3) (denoted by "—"). (d) The unconstrained estimates logit (y_i/n_i) (denoted by "•") of the mother-age-specific logit of Down syndrome incidence, the MLEs $\text{logit}(\widehat{p_i})$ (denoted by "…") subject to the convex constraints via the fastest QLB algorithm, and the MLEs $\text{logit}(\widehat{p_i})$ (denoted by "—") via the pseudo-Newton algorithm.

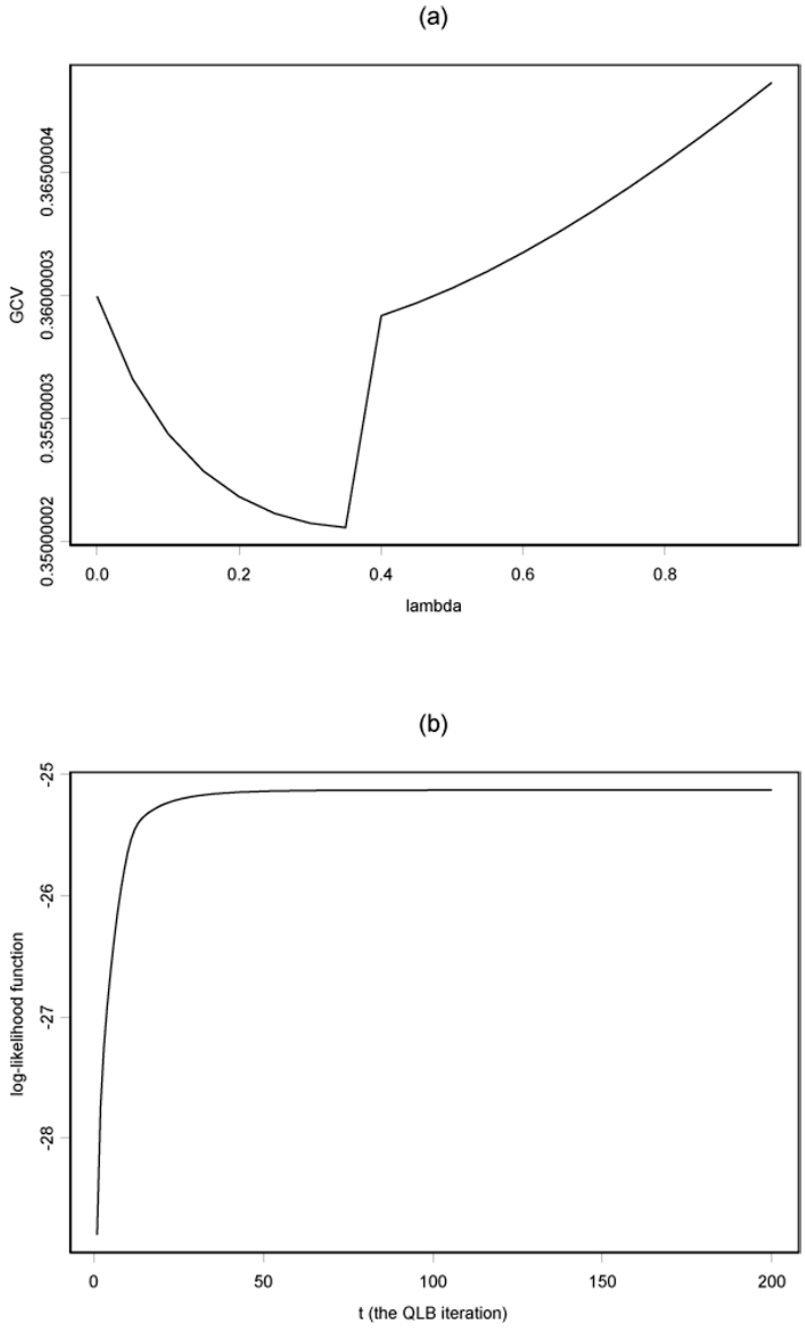

Figure 3.
(a) Plot of generalized cross-validation for the kyphosis data. (b) The monotone convergence of the fastest QLB algorithm (6.3) and (6.4) for the kyphosis data.

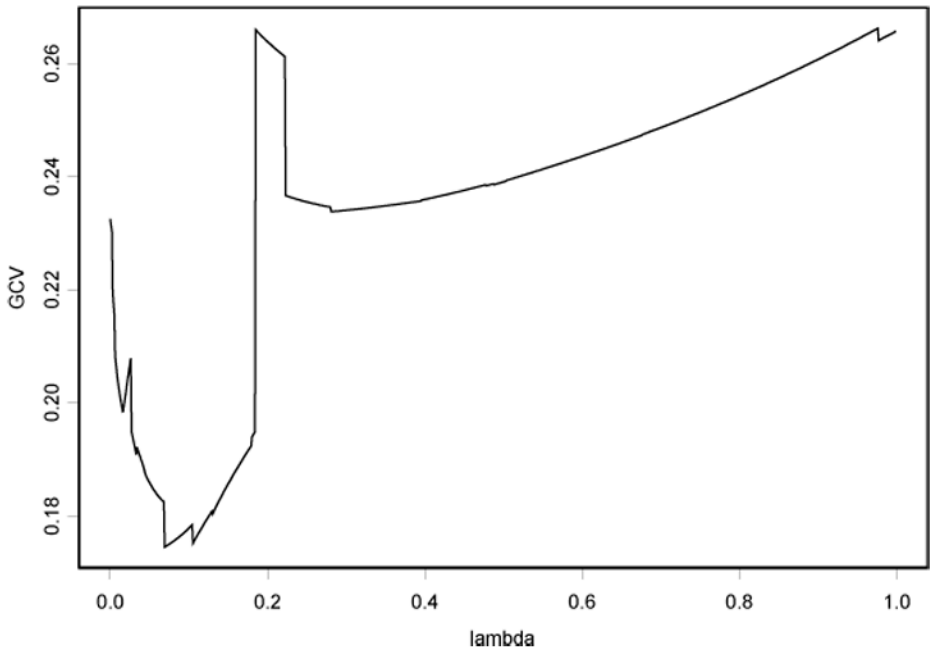

Figure 4.
Plot of generalized cross-validation for the colon microarray data.

Table 1

Simulated data and estimated results for binomial model with umbrella constraints

i	n_i	p_i	y_i	$\hat{p}_i^{\dagger}$	std† of $\hat{p}_i$	i	n_i	p_i	y_i	$\hat{p}_i^{\dagger}$	std† of $\hat{p}_i$
1	5	0.105	1	0.20000	0.06046	21	25	0.705	16	0.63030	0.04806
2	5	0.124	1	0.20000	0.06145	22	25	0.718	14	0.63030	0.06220
3	5	0.133	1	0.20000	0.06408	23	25	0.751	21	0.84000	0.09546
4	5	0.175	1	0.20000	0.06548	24	25	0.771	21	0.84000	0.08251
5	5	0.202	1	0.20000	0.06385	25	25	0.851	22	0.88000	0.07674
6	10	0.215	3	0.30000	0.06566	26	30	0.765	23	0.76667	0.05922
7	10	0.225	4	0.33333	0.05966	27	30	0.724	21	0.70000	0.04463
8	10	0.267	3	0.33333	0.04683	28	30	0.721	19	0.65000	0.05373
9	10	0.289	3	0.33333	0.03551	29	30	0.653	20	0.65000	0.05284
10	10	0.332	5	0.40000	0.04618	30	30	0.612	16	0.53333	0.05790
11	15	0.356	5	0.40000	0.03925	31	35	0.593	18	0.51429	0.04992
12	15	0.478	7	0.40000	0.05279	32	35	0.549	17	0.51429	0.05023
13	15	0.482	5	0.40000	0.05847	33	35	0.442	19	0.51429	0.06220
14	15	0.488	6	0.40000	0.06828	34	35	0.409	10	0.31429	0.06638
15	15	0.501	11	0.63030	0.06134	35	35	0.391	12	0.31429	0.05912
16	20	0.579	12	0.63030	0.05235	36	40	0.343	10	0.25035	0.05917
17	20	0.582	14	0.63030	0.06312	37	40	0.242	9	0.22735	0.04179
18	20	0.616	11	0.63030	0.06312	38	40	0.201	8	0.21268	0.03650
19	20	0.635	17	0.63030	0.05826	38	40	0.156	9	0.20780	0.03606
20	20	0.646	9	0.63030	0.04761	40	40	0.121	1	0.02500	0.00155

$^{\dagger}\{\widehat{p}_i\}_{i=1}^{40}$ and their standard errors were calculated by the fastest QLB algorithm based on the smallest GM matrix B given in (5.2).

Comput Stat Data Anal. Author manuscript; available in PMC 2009 March 15.

Table 2

The Down syndrome data from Hook and Fabia (1978) and estimated results via the pseudo-Newton algorithm

i	a_i	z_i	n_i	y_i	$\hat{p}_i \times 10^4$	std $\times 10^5$	i	a_i	z_i	n_i	y_i	$\hat{p}_i$	std $\times 10^4$
1	15	15.562	1364	1	6.9768	4.7378	18	32	32.479	32116	39	1.261×10^{-3}	0.8936
2	16	16.543	3959	2	6.9768	4.6369	19	33	33.478	28767	47	1.597×10^{-3}	1.1040
3	17	17.527	9848	10	6.9768	4.3887	20	34	34.476	25867	52	2.023×10^{-3}	1.2043
4	18	18.514	19632	9	6.9768	4.1055	21	35	35.474	22947	54	2.562×10^{-3}	1.4869
5	19	19.502	32687	24	6.9768	3.8277	22	36	36.472	19605	70	3.244×10^{-3}	1.7336
6	20	20.493	44376	26	6.9768	3.5367	23	37	37.469	16707	72	4.105×10^{-3}	2.0829
7	21	21.486	51875	30	7.0028	3.3139	24	38	38.466	14006	68	5.195×10^{-3}	2.5338
8	22	22.480	54748	37	7.0359	3.1610	25	39	39.463	10986	50	6.647×10^{-3}	3.2008
9	23	23.475	55757	46	7.0693	3.0853	26	40	40.459	8586	96	8.567×10^{-3}	4.4170
10	24	24.472	54335	34	7.1028	3.1434	27	41	41.454	5729	74	1.103×10^{-2}	6.6370
11	25	25.474	52898	31	7.1367	3.3268	28	42	42.449	3961	51	1.419×10^{-2}	10.314
12	26	26.475	50181	30	7.5825	3.5680	29	43	43.443	2357	44	1.824×10^{-2}	16.555
13	27	27.477	47562	38	8.0567	4.1384	30	44	44.438	1248	23	2.343×10^{-2}	26.750
14	28	28.478	44739	46	8.5600	4.6700	31	45	45.431	638	20	3.003×10^{-2}	42.710
15	29	29.479	41901	42	9.0948	5.2621	32	46	46.425	258	9	3.842×10^{-2}	67.763
16	30	30.480	38106	30	9.6628	6.5066	33	47	47.419	103	7	4.904×10^{-2}	109.73
17	31	31.480	34408	33	10.265	7.8581	34	48	48.411	41	2	6.238×10^{-2}	183.90
							35	49	49.410	13	0	7.917×10^{-2}	496.48

NOTE: i = index, a_i = maternal age in year, z_i = average maternal age, n_i is the number of live births and y_i is the number of Down syndrome cases.

Table 3

Performance of the pseudo-Newton algorithm for the Down syndrome data of Hook and Fabia (1978)

Iteration	$\ell(\theta^{(t)})$	Iteration	$\ell(\theta^{(t)})$	Iteration	$\ell(\theta^{(t)})$
1	−4311.766	6	−797.641	11	−104.273
2	−3336.947	7	−462.382	12	−104.069
3	−2553.834	8	−253.275	13	−104.071
4	−1865.060	9	−142.837	14	−104.069
5	−1273.064	10	−108.280	15	−104.069