Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics and Data Analysis 52 (2008) 3528–3542

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

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

Guo-Liang Tian[a,*], Man-Lai Tang[b], Hong-Bin Fang[a], Ming Tan[a]

[a] *Division of Biostatistics, University of Maryland Greenebaum Cancer Center, 10 South Pine Street, MSTF Suite 261, Baltimore, MD 21201, USA*

[b] *Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China*

Received 5 June 2006; received in revised form 13 November 2007; accepted 14 November 2007
Available online 22 November 2007

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 the 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.

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 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,

[*] Corresponding author. Tel.: +1 410 706 8517; fax: +1 410 706 8548.
E-mail address: gtian2@umm.edu (G.-L. Tian).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.11.007

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

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

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), 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

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

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,

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

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 Sections 5.3 and 5.4). When $\text{rank}(P) = q$, letting $\mu = P\theta$ yields $a \le \mu \le b$ and $\theta = (P^{\top}P)^{-1}P^{\top}\mu$. 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

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

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

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

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

where $J_1(\theta)$ is a penalty function and $\lambda > 0$ is a smoothing parameter for the trade-off 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