Statistics and Its Interface Volume 2 (2009) 493–502

Regularized (bridge) logistic regression for variable selection based on ROC criterion

Guo-Liang Tian*, Hong-Bin Fang, Zhenqiu Liu and Ming T. Tan

It is well known that the bridge regression (with tuning parameter less or equal to 1) gives asymptotically unbiased estimates of the nonzero regression parameters while shrinking smaller regression parameters to zero to achieve variable selection. Despite advances in the last several decades in developing such regularized regression models, issues regarding the choice of penalty parameter and the computational methods for models fitting with parameter constraints even for bridge linear regression are still not resolved. In this article, we first propose a new criterion based on an area under the *receiver operating characteristic* (ROC) curve (AUC) to choose the appropriate penalty parameter as opposed to the conventional generalized cross–validation criterion. The model selected by the AUC criterion is shown to have better predictive accuracy while achieving sparsity simultaneously. We then approach the problem from a constrained parameter model and develop a fast *minorization-maximization* (MM) algorithm for non-linear optimization with positivity constraints for model fitting. This algorithm is further applied to bridge regression where the regression coefficients are constrained with ℓ_p-norm with the level of p selected by data for binary responses. Examples of prognostic factors and gene selection are presented to illustrate the proposed method.

Keywords and phrases: AUC, EM algorithm, Lasso regression, Logistic regression, MM algorithm, ROC, Variable/feature selection.

1. INTRODUCTION

Variable/feature 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. For example, subset selection using the C_p criterion (Mallows, 1973) becomes computationally prohibitive when the number of variables is greater than 50. The forward selection (or forward stepwise regression) is too aggressive (greedy) a fitting technique in that it eliminates at the second step any useful predictors that are correlated with the first selected predictor. A major drawback of the classic methods is that they are numerically unstable in that small changes in data may result in one variable (e.g., a gene) to be selected instead of another due to collinearity. In high dimensional data, the common problem is overfitting. It has been recognized that an effective method to mitigate overfitting and numerical instability is to constrain model parameters, namely, using a regularized regression model such as the lasso regression (Tibshirani, 1996). In spite of advances in developing such regularized regression models, issues regarding the choice of penalty parameter and the computational methods for model fitting with parameter constraints even for bridge linear regression are still not resolved (Wahba, 2007). This article proposes a new criterion for selection of the penalty parameter and an algorithm to fit a special class of regularized regression, i.e., the bridge logistic regression.

Let Y_i denote true disease status of subject i ($Y_i = 1$ if subject i is diseased and $Y_i = 0$ if non-diseased) for $i = 1, 2, \ldots, m$. Let $x_{(i)}$ denote the q-dimensional vector of covariates associated with subject i and θ be a q-dimensional vector of unknown coefficients. We consider the following logistic model,

$$\pi_i = \Pr\{Y_i = 1\} = \frac{\exp(x_{(i)}^\top \theta)}{1 + \exp(x_{(i)}^\top \theta)}, \quad 1 \le i \le m. \tag{1.1}$$

Let Y_i follow the Bernoulli distribution with parameter π_i and y_i denote the realized value of Y_i, then the log-likelihood function is given by

$$L(\theta) = \sum_{i=1}^{m} \{y_i (x_{(i)}^\top \theta) - \log[1 + \exp(x_{(i)}^\top \theta)]\}. \tag{1.2}$$

To formulate the problem, consider the bridge logistic regression with ℓ_p-norm constraint (Frank & Friedman, 1993) which maximizes

$$L(\theta) \quad \text{subject to} \quad \textstyle\sum_{j=1}^{q} |\theta_j|^p \le s, \tag{1.3}$$

where $s\,(> 0)$ is a tuning parameter and $p\,(> 0)$ is a power parameter, with $p = 2$ being the ridge logistic regression, and $p = 1$ being the lasso logistic regression. To maximize (1.3) is equivalent to maximize $L(\theta) - \lambda \sum_{j=1}^{q} |\theta_j|^p$ for $\lambda \ge 0$. That is, for a given $\lambda \in [0, +\infty)$ there exists a $s \ge 0$, such

*Corresponding author.

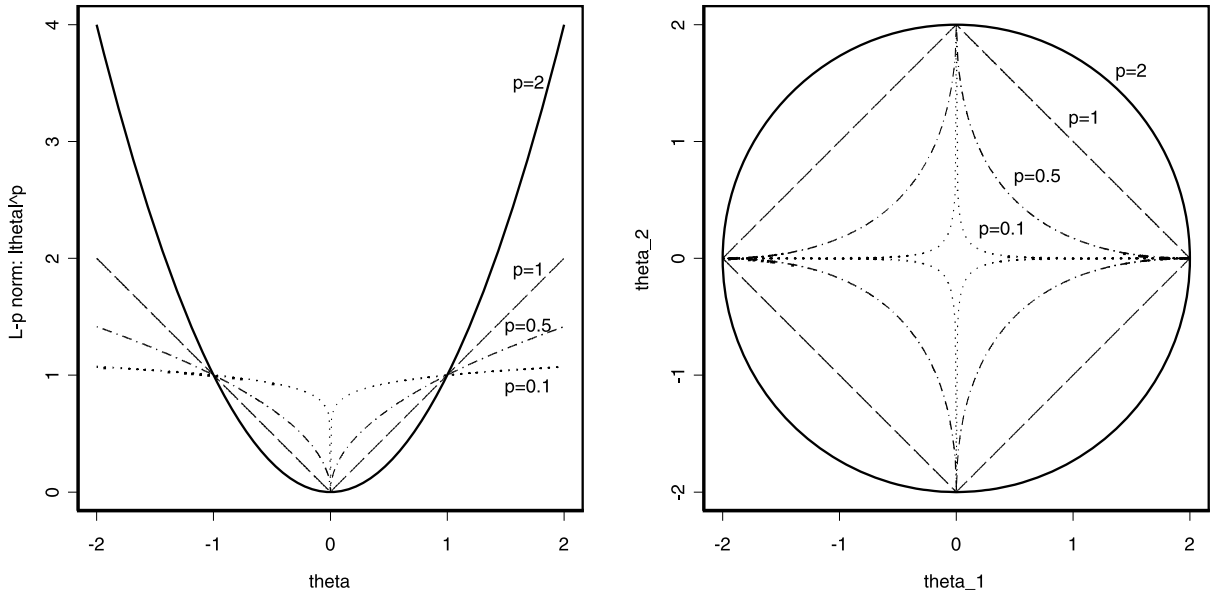

Figure 1. One- and two-dimensional plots for ℓ_p-norm with $p = 0.1, 0.5, 1, 2$.

that the two procedures share the same solution, and vice versa. The goal is to find the penalized *maximum likelihood estimate* (MLE)

(1.4)
$$\hat{\theta}^{\text{bridge}} = \arg\max L_{\lambda,p}(\theta) = \arg\max \{L(\theta) - \lambda \textstyle\sum_{j=1}^{q} |\theta_j|^p\},$$

where $\lambda > 0$ is a penalty parameter. Figure 1 gives one- and two-dimensional plots of ℓ_p-norm for various p values.

Bridge linear regressions with different values of p have very different properties for prediction and classification and have been studied theoretically by Knight & Fu (2000). When $p > 2$, it is shown that the amount of shrinking towards 0 increases with the magnitude of the parameter being estimated and thus for parameters with large values the bias of their estimators may be unacceptably large. When $p \leq 1$, both $\hat{\theta}^{\text{lasso}}$ and $\hat{\theta}^{\text{bridge}}$ regression share the same attractive feature of sparsity, resulting in smaller regression coefficients being 0 (thus selecting variables) if λ is sufficiently large. Thus, the method combines parameter estimation and variable selection. However, the applications of lasso ($p = 1$) and its variant with *smoothly clipped absolute deviation* (SCAD) (Fan & Li, 2001, 2002) regression have been limited for several reasons. First, the original lasso algorithm involves an iterative step within each reweighted least square and may converge slowly or not converging at all (Tibshirani, 1996). The method becomes highly inefficient when the number of covariates q is large. As pointed out by Madigan & Ridgeway (2004), the relative inefficiency of the original lasso algorithm and the relative complexity of more recent lasso algorithms (e.g., Osborne et al., 2000) may be to blame.

Motivated in part by improving the slow convergence of lasso, Efron et al. (2004) proposed the *least angle regressions* (LARS) as a new variable selection procedure, which, in fact, leads to lasso. However, the method requires a search stopping rule, which is currently available only for linear regression, and LARS may also lead to overfitting (Stine, 2004). In addition, an extension of LARS-type strategies to generalized linear models encounters greater computational challenges such as nonlinear optimization (Madigan & Ridgeway, 2004), in particular, the ℓ_1-constrained solution in logistic regression is not piecewise linear and hence the pathwise optimization is more difficult (Efron et al., 2004, p. 497). In contrast, lasso/bridge regression requires no stopping rule as in stepwise regression, and it builds on the simple idea of regression with the ℓ_p-penalty. However, when $p < 1$, the bridge regression gives asymptotically unbiased estimates of the nonzero regression parameters consistently while shrinking the estimates of zero (or small) regression parameters to zero (Knight & Fu, 2000), implying potentially better predictive performance (Malioutov et al., 2005). Unfortunately, Frank & Friedman (1993) do not provide computational method for bridge linear regression for any given λ and p and the method proposed by Fu (1998) is available only for $p > 1$. Recently, we proposed an approximate solution by using a smoothed penalty function $(\theta_i^2 + \epsilon)^{p/2}$ which approaches to ℓ_p-penalty ($p < 1$) when $\epsilon \to 0$ (Liu et al., 2007). This approximate approach employed another parameter ϵ whose value has to be pre-specified and its accuracy is not yet quantified. Therefore, the major hurdle in bridge regression continues to be computational.

The EM-type algorithms have emerged as a powerful tool for optimization with *linear inequality constraints* (LICs)

(Liu, 2000; Tan, Tian & Fang, 2003). Recently, Tan et al. (2007) developed a fast EM algorithm for quadratic optimization subject to box constraints and LICs, which provides a promising algorithm to bridge linear regression with $p < 1$. In addition, in the most existing methods, penalty parameters are selected by minimizing the approximate *generalized cross-validation* (GCV) statistic (Craven & Wahba, 1979). The GCV in this case is based on both sensitivity and specificity of the predictive model. This method is not optimal in biomedical applications since often the numbers of normal and cancer specimens are different and the sensitivity and the specificity depend on the cutoff point chosen to derive the predictive (cancer or non-cancer) rule. The area under the *receiver operating characteristic* (ROC) is known to be a better measure for predictive power. Thus, an optimal penalty parameter selected via maximizing the area under the ROC curve (AUC) statistic is more desirable.

Therefore, the purpose of this article is to introduce an efficient alternative model fitting method and to utilize the AUC to choose the appropriate penalty parameter in the bridge regression model as opposed to the conventional GCV criterion. §2 provides automatic selection of the penalty and power parameters $\hat{\lambda}^{\text{opt}}$ and $\hat{p}^{\text{opt}}$ via maximizing the AUC statistic instead of minimizing an approximate GCV statistic. We then develop a fast *minorization–maximization* (MM) algorithm for non-linear optimization with positivity constraints for model fitting in §3. This algorithm is then applied to bridge logistic regression with $p \le 1$. As in the lasso linear regression, the unconstrained MLEs of regression coefficients are used as the initial values, thus the proposed algorithm can only deal with the problems where the number of covariates is less than the sample size (i.e., $q < m$). Examples of prognostic factors and micro–array analysis are presented in §4. We conclude with a discussion.

2. DATA-DRIVEN CHOICE OF THE PENALTY AND POWER PARAMETERS VIA THE AUC CRITERION

2.1 The ROC curve

For given $\lambda > 0$ and $0 < p \le 1$, we calculate the bridge estimate from (1.4) and denote it by $\hat{\theta}^{\text{bridge}}_{\lambda,p}$, which depends on both λ and p. Therefore, for a given covariate x, the prediction probability is given by

$$\hat{\pi}_{\lambda,p} = \widehat{\Pr}\{Y = 1\} = \frac{\exp(x^\top \hat{\theta}^{\text{bridge}}_{\lambda,p})}{1 + \exp(x^\top \hat{\theta}^{\text{bridge}}_{\lambda,p})}. \tag{2.1}$$

With a threshold $c \in (0, 1)$, we define a binary test T as follows:

$$\begin{aligned} T = + \quad &\text{if} \quad \hat{\pi}_{\lambda,p} \ge c, \\ T = - \quad &\text{if} \quad \hat{\pi}_{\lambda,p} < c. \end{aligned}$$

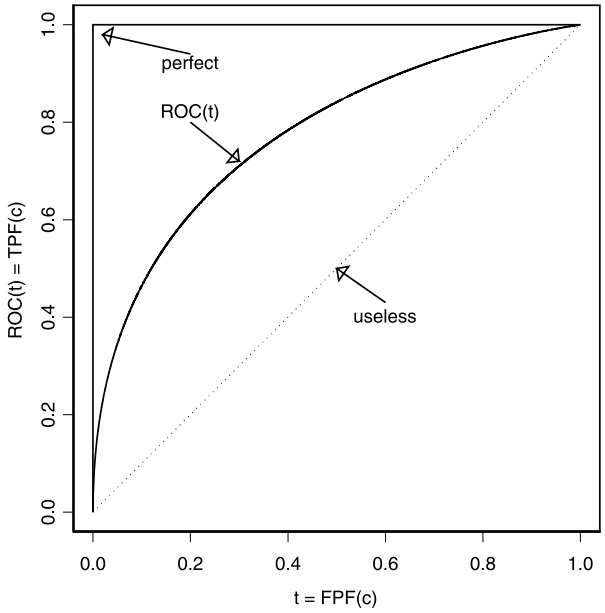

Figure 2. ROC curves for a useless test/prediction and a perfect test/prediction.

Furthermore, let

$$\begin{aligned} &\text{FPF}_{\lambda,p}(c) = \Pr\{T = +|Y = 0\} \quad \text{and} \\ &\text{TPF}_{\lambda,p}(c) = \Pr\{T = +|Y = 1\} \end{aligned} \tag{2.2}$$

denote false and true positive fractions at the threshold c, respectively, then, the ROC is defined as (e.g., see, Pepe, 2003, p. 67–68)

$$\begin{aligned} \text{ROC}_{\lambda,p}(\cdot) &= \{(\text{FPF}_{\lambda,p}(c), \text{TPF}_{\lambda,p}(c)),\ c \in (0,1)\} \\ &= \{(t, \text{ROC}_{\lambda,p}(t)),\ t \in (0,1)\}. \end{aligned} \tag{2.3}$$

It has been shown that the ROC curve is a monotone increasing function mapping $(0, 1)$ onto $(0, 1)$. A useless test/prediction (corresponding to a poor choice of λ and p) is one such that the distribution functions for T are the same in the diseased and non-diseased populations. The ROC curve for a useless test/prediction is then $\text{ROC}_{\lambda,p}(t) = t$. On the other hand, a perfect test/prediction (corresponding to a good choice of λ and p) entirely separates diseased and non-diseased subjects. Its ROC curve is along the left and upper borders of the first unit quadrant. Better tests/predictions have ROC curves closer to the upper left corner. These are illustrated in Figure 2.

In literature, several numerical indices are proposed to summarize ROC curves. The most commonly used summary measure is the AUC, which is defined as

$$\text{AUC}(\lambda, p) = \int_0^1 \text{ROC}_{\lambda,p}(t)\, dt. \tag{2.4}$$

We determine the optimal $(\hat{\lambda}^{\text{opt}}, \hat{p}^{\text{opt}})$ by maximizing $\text{AUC}(\lambda, p)$ over a grid of $\lambda > 0$ and $0 < p \le 1$.

2.2 Empirical estimation of the AUC

Let $Y_{\text{obs}} = \{(x_{(1)}, y_1), \ldots, (x_{(m)}, y_m)\}$ denote m observation data, where $x_{(i)}$ is the q-dimensional covariate vector for subject i and $y_i \in \{0, 1\}$. For a given pair of (λ, p), we optimize (1.4) and obtain $\hat{\theta}^{\text{bridge}}_{\lambda,p}$. For each subject, based on (2.1), we then calculate the m prediction probabilities:

$$\hat{\pi}_{i(\lambda,p)} = \widehat{\Pr}\{Y_i = 1\} = \frac{\exp(x_{(i)}^\top \hat{\theta}^{\text{bridge}}_{\lambda,p})}{1 + \exp(x_{(i)}^\top \hat{\theta}^{\text{bridge}}_{\lambda,p})}, \quad i = 1, \ldots, m. \tag{2.5}$$

Without loss of the generality, we assume that the first m_0 subjects are non-diseased (or controls) and the rest $m_1 = m - m_0$ subjects are diseased (or cases). Thus, for each cut–point $c \in (0, 1)$, the false and true positive fractions in (2.2) are estimated by

$$\begin{aligned} \widehat{\text{FPF}}_{\lambda,p}(c) &= \frac{1}{m_0} \sum_{j=1}^{m_0} I(\hat{\pi}_{j(\lambda,p)} \ge c) \qquad \text{and} \\ \widehat{\text{TPF}}_{\lambda,p}(c) &= \frac{1}{m_1} \sum_{i=m_0+1}^{m} I(\hat{\pi}_{i(\lambda,p)} \ge c), \end{aligned} \tag{2.6}$$

respectively, where $I(\cdot)$ denotes the indicator function. The estimated ROC curve, denoted by $\widehat{\text{ROC}}_{\lambda,p}(t)$, is a plot of $\widehat{\text{TPF}}_{\lambda,p}(c)$ versus $\widehat{\text{FPF}}_{\lambda,p}(c)$ for all $c \in (0, 1)$. In addition, it has been shown that the estimated AUC is given by (e.g., see, Pepe, 2003, p.103–104)

$$\begin{aligned} \widehat{\text{AUC}}(\lambda, p) = \int_0^1 \widehat{\text{ROC}}_{\lambda,p}(t)\, dt = \frac{1}{m_0 m_1} \sum_{j=1}^{m_0} \sum_{i=m_0+1}^{m} \\ \times \left\{ I(\hat{\pi}_{i(\lambda,p)} > \hat{\pi}_{j(\lambda,p)}) + \frac{1}{2} I(\hat{\pi}_{i(\lambda,p)} = \hat{\pi}_{j(\lambda,p)}) \right\}, \end{aligned} \tag{2.7}$$

which is exactly the Wilcoxon or Mann–Whitney U-statistic.

3. AN MM ALGORITHM WITH MONOTONIC CONVERGENCE

3.1 Formulation of the algorithm

Let $\theta^{(t)}$ denote the current approximation of $\hat{\theta}^{\text{bridge}}$ defined in (1.4). For a given $\theta^{(t)}$, $Q_{\lambda,p}(\theta|\theta^{(t)})$ is a real–valued function depending on both (λ, p). The function $Q_{\lambda,p}(\theta|\theta^{(t)})$ is said to minorize $L_{\lambda,p}(\theta)$ at $\theta^{(t)}$ if

$$Q_{\lambda,p}(\theta|\theta^{(t)}) \le L_{\lambda,p}(\theta) \qquad \text{for all } \theta, \tag{3.1}$$

$$Q_{\lambda,p}(\theta^{(t)}|\theta^{(t)}) = L_{\lambda,p}(\theta^{(t)}). \tag{3.2}$$

With the MM algorithm (Lange et al., 2000), we maximize the minorizing function $Q_{\lambda,p}(\theta|\theta^{(t)})$ instead of the target function $L_{\lambda,p}(\theta)$. If $\theta^{(t+1)}$ is the maximizer of $Q_{\lambda,p}(\theta|\theta^{(t)})$, i.e.,

$$\theta^{(t+1)} = \arg\max Q_{\lambda,p}(\theta|\theta^{(t)}), \tag{3.3}$$

then, from (3.1) and (3.2), we have

$$L_{\lambda,p}(\theta^{(t+1)}) \ge Q_{\lambda,p}(\theta^{(t+1)}|\theta^{(t)}) \ge Q_{\lambda,p}(\theta^{(t)}|\theta^{(t)}) = L_{\lambda,p}(\theta^{(t)}). \tag{3.4}$$

Under appropriate additional compactness and continuity conditions, the ascent property (3.4) guarantees the monotone convergence of the MM algorithm (De Leeuw, 2006). From (3.4) we can see that it is not necessary to actually maximize the minorizing function, it suffices to find $\theta^{(t+1)}$ such that $Q_{\lambda,p}(\theta^{(t+1)}|\theta^{(t)}) \ge Q_{\lambda,p}(\theta^{(t)}|\theta^{(t)})$.

3.2 The sharpest quadratic minorizing function

Let ∇ denote the derivative operator. From (1.2), the score vector and the observed information matrix are given by

$$\begin{aligned} \nabla L(\theta) &= \textstyle\sum_{i=1}^m (y_i - \pi_i) x_{(i)} = X^\top (y - \pi), \\ -\nabla^2 L(\theta) &= \textstyle\sum_{i=1}^m \pi_i (1 - \pi_i) x_{(i)} x_{(i)}^\top = X^\top D X, \end{aligned}$$

respectively, where $X^\top = (x_{(1)}, \ldots, x_{(m)})$, $y = (y_1, \ldots, y_m)^\top$, $\pi = (\pi_1, \ldots, \pi_m)^\top$, and

$$D = \text{diag}(\pi_1(1 - \pi_1), \ldots, \pi_m(1 - \pi_m)). \tag{3.5}$$

Since $\pi_i(1 - \pi_i) \le 1/4$ for $\pi_i \in [0, 1]$, $B \hat{=} (1/4) X^\top X$ is a positive definite matrix and globally majorizes the observed information, i.e., $B \ge -\nabla^2 L(\theta)$ for all θ. Therefore,

$$\begin{aligned} Q^{\text{B}}_{\lambda,p}(\theta|\theta^{(t)}) = L(\theta^{(t)}) + (\theta - \theta^{(t)})^\top \nabla L(\theta^{(t)}) \\ - 0.5(\theta - \theta^{(t)})^\top B (\theta - \theta^{(t)}) - \lambda \textstyle\sum_{j=1}^q |\theta_j|^p \end{aligned} \tag{3.6}$$

minorizes $L_{\lambda,p}(\theta)$ at $\theta^{(t)}$ (Böhning & Lindsay, 1988). However, this minorizing function is not very sharp. The sharpest minorizing function, discovered independently by Jaakkola & Jordan (2000) and Groenen et al. (2003), is given by

$$Q^{\text{S}}_{\lambda,p}(\theta|\theta^{(t)}) = Q^{\text{S}}(\theta|\theta^{(t)}) - \lambda \sum_{j=1}^q |\theta_j|^p, \tag{3.7}$$

where

$$\begin{aligned} Q^{\text{S}}(\theta|\theta^{(t)}) &\hat{=} L(\theta^{(t)}) + (\theta - \theta^{(t)})^\top \nabla L(\theta^{(t)}) \\ &\quad - 0.5(\theta - \theta^{(t)})^\top S(\theta^{(t)})(\theta - \theta^{(t)}), \\ S(\theta) &\hat{=} X^\top W(\theta) X, \\ W(\theta) &= \text{diag}((\pi_1 - 0.5)/x_{(1)}^\top\theta, \ldots, (\pi_m - 0.5)/x_{(m)}^\top\theta). \end{aligned}$$

Furthermore, we have

$$Q^{\mathrm{B}}_{\lambda,p}(\theta|\theta^{(t)}) \le Q^{\mathrm{S}}_{\lambda,p}(\theta|\theta^{(t)}) \le L_{\lambda,p}(\theta) \qquad \text{for all } \theta,$$
$$Q^{\mathrm{B}}_{\lambda,p}(\theta^{(t)}|\theta^{(t)}) = Q^{\mathrm{S}}_{\lambda,p}(\theta^{(t)}|\theta^{(t)}) = L_{\lambda,p}(\theta^{(t)}).$$

3.3 Quadratic optimization with positivity constraints

The MM algorithm can be applied to obtain $\hat{\theta}^{\mathrm{bridge}}$ by iteratively computing

$$\theta^{(t+1)} = \arg\min\{-Q^{\mathrm{S}}(\theta|\theta^{(t)}) + \lambda \textstyle\sum_{j=1}^{q} |\theta_j|^p\}. \tag{3.8}$$

For $m > q$, let $\hat{\theta}^{\mathrm{U}}$ denote the unconstrained MLE of θ in the logistic model (1.2) and $v = (v_1, \ldots, v_q)^\top$ be its sign vector (i.e., $v_j = \mathrm{sign}(\hat{\theta}^{\mathrm{U}}_j) = +1, 0$, or -1 corresponding to positive, zero, or negative values of $\hat{\theta}^{\mathrm{U}}_j$). We first show that both $\hat{\theta}^{\mathrm{U}}$ and $\hat{\theta}^{\mathrm{bridge}}$ have the same signs.

For the normal linear regression where the likelihood $L(\theta)$ is a quadratic function of θ, Lemma 7 and Lemma 8 of Efron et al. (2004) proved that both the unconstrained MLE $\hat{\theta}^{\mathrm{U}}$ and $\hat{\theta}^{\mathrm{lasso}}$ share signs. The geometric shape of the second plot of Figure 1 (or see Figure 2 of Tibshirani, 1996) further shows that $\hat{\theta}^{\mathrm{U}}$, $\hat{\theta}^{\mathrm{ridge}}$, $\hat{\theta}^{\mathrm{lasso}}$, and $\hat{\theta}^{\mathrm{bridge}}$ have the same signs for the case of normal linear regression.

For the logistic regression, although the likelihood $L(\theta)$ specified by (1.2) is a non-linear function of θ, the MM algorithm (3.8) implies that finding $\hat{\theta}^{\mathrm{bridge}}$ is equivalent to iteratively finding the maximizer of the quadratic function $Q^{\mathrm{S}}(\theta|\theta^{(t)})$ with ℓ_p-penalty. In addition, from §3.2, $\hat{\theta}^{\mathrm{U}}$ is also the solution of maximizing $Q^{\mathrm{S}}(\theta|\theta^{(t)})$ as $t \to \infty$. Thus, both $\hat{\theta}^{\mathrm{bridge}}$ and $\hat{\theta}^{\mathrm{U}}$ share signs for the case of logistic regression. This implies the bridge estimator $\hat{\theta}^{\mathrm{bridge}} \in \{(v_1\beta_1, \ldots, v_q\beta_q)^\top = \mathrm{diag}(v)\beta : \ \beta \in \mathbb{R}^q_+\}$. Given $\theta^{(t)}$ and from (3.8), we have

$$\theta^{(t+1)} = \mathrm{diag}(v) \cdot \beta^{(t+1)}, \tag{3.9}$$
$$\beta^{(t+1)} = \arg\min_{\beta \in \mathbb{R}^q_+} \{-Q^{\mathrm{S}}(\mathrm{diag}(v)\beta|\theta^{(t)}) + \lambda \textstyle\sum_{j=1}^{q} \beta_j^p\}. \tag{3.10}$$

The built-in S-Plus function `nlminb` (nonlinear minimization subject to box constraints) can applied to (3.10). Especially, when $p = 1$, the target function in (3.10) is a quadratic function, thus, we can utilize the built-in S-PLUS function `nnls.fit` (linear least-squares with nonnegative constraints) to solve (3.10) iteratively.

3.4 Standard errors

With the efficient algorithm developed in §3.3 for computing $\hat{\theta}^{\mathrm{bridge}}$, calculating the standard errors of $\hat{\theta}^{\mathrm{bridge}}$ via bootstrapping (Efron & Tibshirani, 1993) becomes computationally feasible. Having obtained the $\hat{\theta}^{\mathrm{bridge}}$ based on (3.9) and (3.10), we can directly generate a bootstrap sample $\{y_i^*\}_{i=1}^m$ with

$$y_i^* \overset{\text{ind}}{\sim} \text{Bernoulli}\Big(\exp\{x_{(i)}^\top \hat{\theta}^{\mathrm{bridge}}\}/[1 + \exp\{x_{(i)}^\top \hat{\theta}^{\mathrm{bridge}}\}]\Big)$$

and compute the corresponding bootstrap replication $\hat{\theta}^*$. Independently repeating this process G times, we obtain G bootstrap replications $\{\hat{\theta}^*(g)\}_{g=1}^G$, where $\hat{\theta}^*(g) = (\hat{\theta}^*_1(g), \ldots, \hat{\theta}^*_q(g))^\top$. Therefore, the standard error $\mathrm{se}(\hat{\theta}^{\mathrm{bridge}}_j)$ of $\hat{\theta}^{\mathrm{bridge}}_j$ can be estimated by the sample standard deviation of the G replications.

4. TWO DATA ANALYSIS EXAMPLES

4.1 Kyphosis data

This data set consists of retrospective measurements on 83 laminectomy patients (Hastie & 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 quadratic terms in the model after centering each of the three variables. For comparison purposes, we did not include the interaction terms. Since all the covariates are continuous, they are standardized individually in our analysis. The full logistic regression model is

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

The SAS `proc logistic` with `backward` stepwise selection removed the x_2^2-term and the resulting estimates of the regression coefficients are listed in the 4-th column of Table 1.

To apply the proposed MM algorithm (3.9) and (3.10) to obtain the bridge solution $\hat{\theta}^{\mathrm{bridge}}$, we first need to calculate the unconstrained MLE. We have

$$\hat{\theta}^{\mathrm{U}} = (-2.6422, 0.8270, 0.7673, -2.2688, -1.5406, 0.0321, -1.1582)^\top$$

and its sign vector $v = (-1, 1, 1, -1, -1, 1, -1)^\top$. The AUC criterion is used to select the optimal penalty and power parameters. We obtain $\hat{\lambda}^{\mathrm{opt}} = 5.41$ and $\hat{p}^{\mathrm{opt}} = 0.1$ (see Figure 3(b)). The resulting $\hat{\theta}^{\mathrm{bridge}}$ and AUC are displayed in the 9-th column of Table 1. The corresponding standard errors with 1,000 bootstrap replications are 0.5181, 0.4287, 0.3934, 0.7612, 0.5623, (−), and 0.4076, respectively. When we fix $p = 1$ and repeat the above process, we obtain $\hat{\lambda}^{\mathrm{opt}} = 0.704$ (see Figure 3(a)). The corresponding lasso solution $\hat{\theta}^{\mathrm{lasso}}$ is given in the 8-th column of Table 1. As expected, the AUC induced by $\hat{\theta}^{\mathrm{lasso}}$ is less than the AUC induced by $\hat{\theta}^{\mathrm{bridge}}$.

To compare the proposed AUC criterion with the existing GCV criterion, for any given $\lambda > 0$ and $0 < p \le 1$, we calculate the bridge estimate from (1.4) and denote it by $\hat{\theta}^{\mathrm{bridge}}_{\lambda,p}$. The GCV statistic is defined as

$$\mathrm{GCV}(\lambda, p) = -L(\hat{\theta}^{\mathrm{bridge}}_{\lambda,p})/\{m[1 - e(\lambda)/m]^2\},$$

Table 1. Comparisons of the bridge and lasso regressions under the GCV and AUC criteria

Variable	Parameter	MLE†	Backward stepwise	Tibshirani's lasso‡	GCV		AUC	
					lasso	bridge	lasso	bridge
Intercept	θ_0	−2.6422	−2.6451	−1.42	−2.1802	−2.0723	−1.8814	−1.9093
x_1	θ_1	0.8270	0.8310	0.03	0.7749	0.6830	0.5452	0.5581
x_2	θ_2	0.7673	0.7955	0.31	0.5690	0.5109	0.4429	0.4460
x_3	θ_3	−2.2688	−2.2670	−0.48	−1.8320	−1.6910	−1.4466	−1.4631
x_1^2	θ_4	−1.5406	−1.5320	−0.28	−1.1462	−1.0636	−0.9158	−0.9406
x_2^2	θ_5	0.0321	0.0000	0.00	0.0000	0.0000	0.0000	0.0000
x_3^2	θ_6	−1.1582	−1.1533	0.00	−0.9054	−0.7971	−0.6205	−0.6153
$\hat{\lambda}^{\text{opt}}$	–	–	–	–	0.085	50.13	0.704	5.41
$\hat{p}^{\text{opt}}$	–	–	–	1	1	0.005	1	0.1
AUC	–	0.9162	0.9170	0.8692	0.9179	0.9171	0.9205	0.9214

†Unconstrained MLE.
‡The results obtained by Tibshirani (1996) and it is not clear which criterion (CV, GCV and Stein unbiased estimate of risk) was used in his paper.

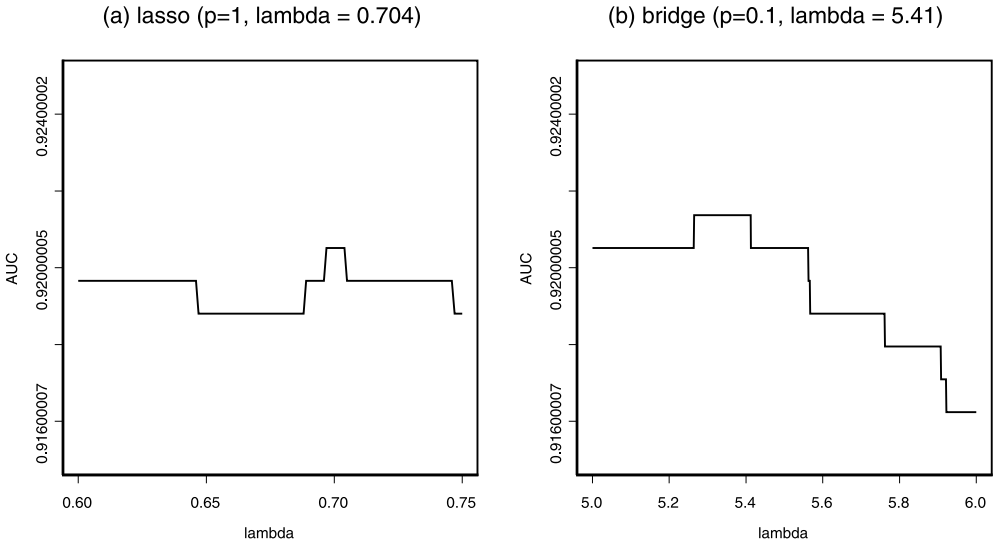

Figure 3. Plot of AUC versus λ for the kyphosis data. (a) Lasso regression with $p = 1$ and $\hat{\lambda}^{\text{opt}} = 0.704$; (b) Bridge regression with $\hat{p}^{\text{opt}} = 0.1$ and $\hat{\lambda}^{\text{opt}} = 5.41$.

where $e(\lambda) = \text{tr}[X(X^\top DX + \lambda W^-)^{-1}X^\top D]$ is the effective number of parameters, W^- enotes the Moore-Penrose generalized inverse of $W = \text{diag}(|\hat{\theta}^{\text{bridge}}_{\lambda,p}|)$, and D is defined by (3.5). We determine the optimal $(\hat{\lambda}^{\text{opt}}, \hat{p}^{\text{opt}})$ by minimizing $\text{GCV}(\lambda, p)$ over a grid of $\lambda > 0$ and $0 < p \leq 1$. Figure 4 shows that the optimal $\hat{\lambda}^{\text{opt}} = 0.085$ for the lasso regression, while the optimal $\hat{p}^{\text{opt}} = 0.005$ and $\hat{\lambda}^{\text{opt}} = 50.13$ for the bridge regression.

Based on the GCV criterion, we obtain the corresponding bridge and lasso estimates (see Table 1). However, the lasso estimates obtained by Tibshirani (1996) are

$$-1.42 + 0.03x_1 + 0.31x_2 - 0.48x_3 - 0.28x_1^2,$$

which differ from ours. To some extent, this is expected. Tibshirani (1996) showed that different criteria (e.g., CV, GCV and Stein unbiased estimate of risk) could result in different choices of the tuning parameter s. It is not clear which one was actually used in the computation from the paper. Figure 5 shows the comparison of ROC curves between the bridge regression under AUC criterion and backward stepwise, Tibshirani's lasso regression, lasso regression under GCV criterion, and bridge regression under GCV criterion. The corresponding AUC values are given in the last row of the Table 1. As expected, the AUC for the bridge regression under AUC criterion is the highest.

4.2 Colon microarray data

The colon microarray data set is composed of 2,000 genes per sample in 22 normal colon tissue samples and 40 tumor colon samples (Alon et al., 1999). The outcome is bi-

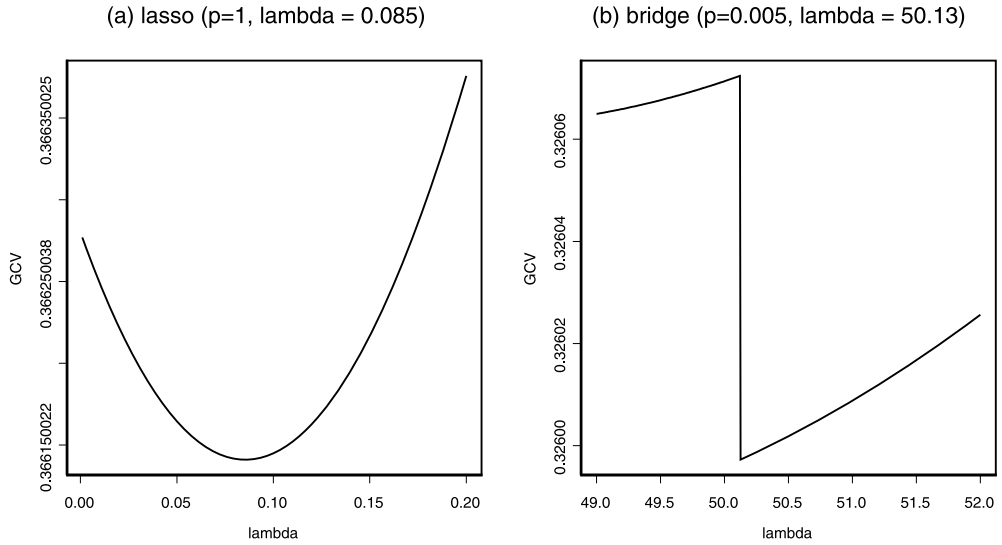

Figure 4. Plot of GCV versus λ for the kyphosis data. (a) Lasso regression with $p = 1$ and $\hat{\lambda}^{\mathrm{opt}} = 0.085$; (b) Bridge regression with $\hat{p}^{\mathrm{opt}} = 0.005$ and $\hat{\lambda}^{\mathrm{opt}} = 50.13$.

Table 2. Computational results for the colon microarray data

Gene ID	Variable	Parameter	$\hat{\theta}^{\mathrm{U}}$	v	$\hat{\theta}^{\mathrm{bridge}}$	
					AUC = 0.98	AUC = 0.852
	Intercept	θ_0	−9.908	−1	−2.052	−4.185
493	x_1	θ_1	3.795	1	0.938	0
1423	x_2	θ_2	18.414	1	2.196	0.408
249	x_3	θ_3	5.191	1	0.517	0
377	x_4	θ_4	−3.800	−1	0	0
765	x_5	θ_5	13.509	1	1.891	0.123
245	x_6	θ_6	−33.823	−1	−2.229	−1.604
267	x_7	θ_7	−3.371	−1	−2.681	−2.376
1635	x_8	θ_8	0.182	1	0	0
66	x_9	θ_9	6.184	1	0.529	0
625	x_{10}	θ_{10}	−9.619	−1	−1.827	−1.533
14	x_{11}	θ_{11}	7.296	1	1.674	0.692
822	x_{12}	θ_{12}	9.060	1	1.056	0
1892	x_{13}	θ_{13}	4.765	1	0.459	0
1494	x_{14}	θ_{14}	−27.769	−1	−2.710	−0.662
137	x_{15}	θ_{15}	4.486	1	0.963	0
897	x_{16}	θ_{16}	−3.973	−1	−0.853	0
111	x_{17}	θ_{17}	10.797	1	0.833	0
513	x_{18}	θ_{18}	−4.303	−1	−1.106	0
1843	x_{19}	θ_{19}	16.403	1	1.792	0
812	x_{20}	θ_{20}	5.354	1	0.990	0
739	x_{21}	θ_{21}	5.136	1	0.459	0
780	x_{22}	θ_{22}	1.811	1	0.932	0
286	x_{23}	θ_{23}	−2.696	−1	−0.700	0
1060	x_{24}	θ_{24}	−1.708	−1	−0.681	−0.483
415	x_{25}	θ_{25}	−17.991	−1	−1.642	−0.435

nary (1 = tumor colon, 0 = normal colon). The data set was first normalized for each gene to have a zero mean and unit variance. For $s = 1, \ldots, 2{,}000$, 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. Only 25 out of 2,000 genes are identified to be marginally significant at the 0.001 level and the corresponding gene IDs are displayed in the first column of Table 2.

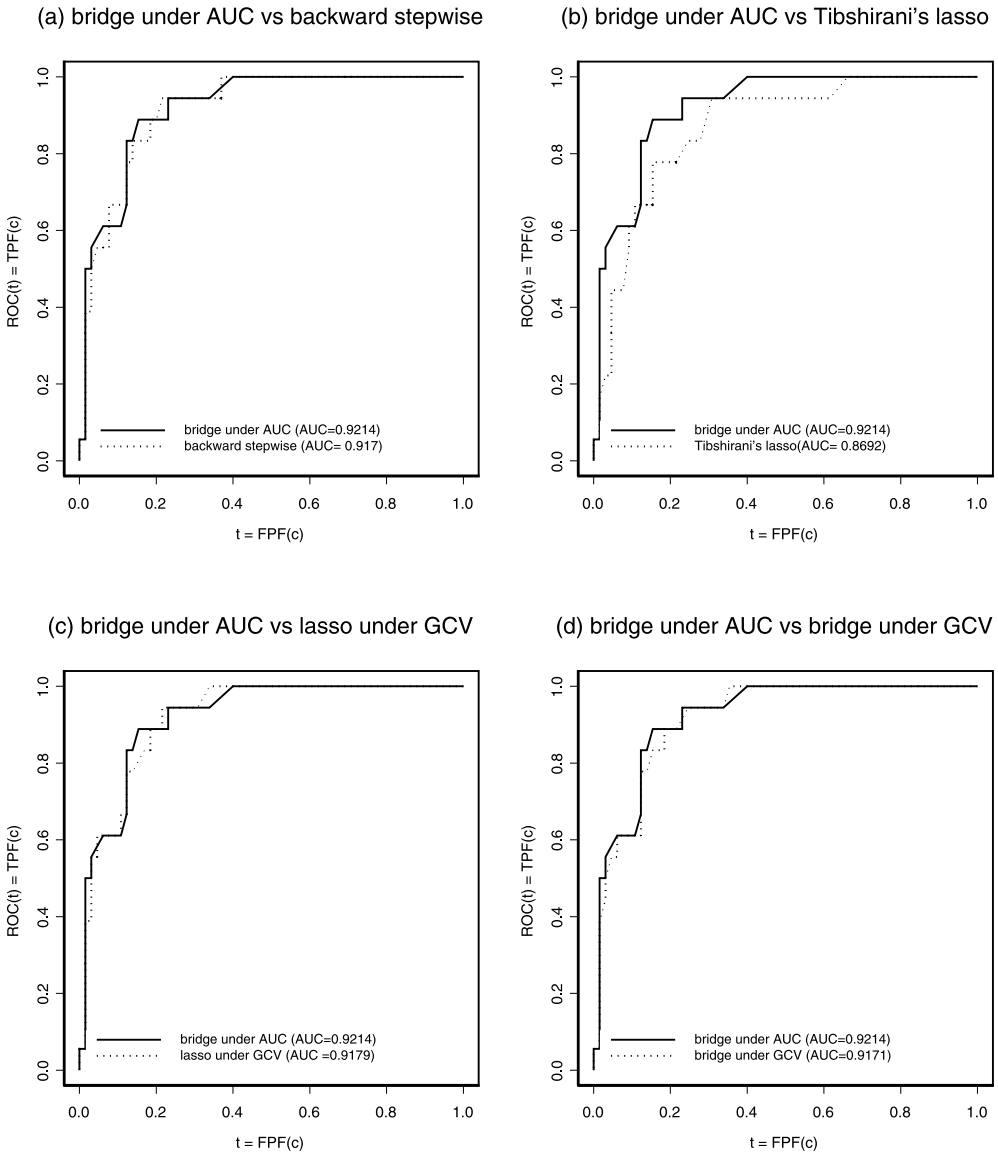

Figure 5. Comparisons of ROC curves for the kyphosis data. (a) Bridge regression under AUC criterion vs. backward stepwise; (b) Bridge regression under AUC criterion vs. Tibshirani's lasso regression; (c) Bridge regression under AUC criterion vs. lasso regression under GCV criterion; (d) Bridge regression under AUC criterion vs. bridge regression under GCV criterion.

We first compute the unconstrained MLE $\hat{\theta}^{\mathrm{U}}$ (4-th column of Table 2). The corresponding sign vector v is given in the 5-th column of Table 2. Under the AUC criterion, the optimal $\hat{\lambda}^{\mathrm{opt}} = 0.95$ and the optimal $\hat{p}^{\mathrm{opt}} = 0.009$. Using $\theta^{(0)} = v$ as the initial values, the proposed MM algorithm (3.9) and (3.10) converged to the bridge estimator $\hat{\theta}^{\mathrm{bridge}}$ (6-th column of Table 2). That is, 23 out of the 25 genes are identified under the AUC criterion. The corresponding AUC is 0.98. Apparently, the larger the AUC is, the more genes selected. If the number of selected genes is less than 10, then $\lambda = 98.3724$ and $p = 0.009$, resulting in 9 genes being selected from the 25 genes. The resulting regression coefficients are given in the last column of Table 2 and the corresponding AUC is 0.852.

5. DISCUSSION

We proposed an alternative regularized (bridge) logistic regression using the AUC criterion instead of the GCV to select the optimal penalty parameter λ and power parameter p because the AUC considers both the sensitivity and specificity. The proposed MM algorithm transfers the original bridge optimization problem (1.4) into a series of simple optimization problems (3.8) by replacing the likelihood function with a quadratic surrogate function. A key step of the fast MM algorithm is to utilize the property that parameter estimates from lasso, bridge, ridge regressions and the unconstrained MLE share signs so that the absolute-value in the penalty function in (3.8) can be removed, resulting in

a series of much simpler optimizations with positivity constraints where the target function in (3.10) is continuous and differentiable everywhere.

Note that the bridge penalty with $0 < p < 1$ in (3.10) is not convex and is singular at zero, its behavior around zero (i.e., small estimated coefficients) may be erratic if the whole target function

$$-Q^{\mathrm{S}}(\mathrm{diag}(v)\beta|\theta^{(t)}) + \lambda\sum_{j=1}^{q}\beta_j^p \tag{5.1}$$

in (3.10) is not convex. In practice, this could be a common issue with any algorithms for the regularized (bridge) logistic regression. However, once (5.1) is convex for some p, the MM algorithm can guarantee monotone convergence. In fact, we did not encounter this kind of unstable phenomenon at least in our two data-analysis examples.

The MM algorithm is preferable when the number of variables is not too large because its stable convergence. Otherwise, we can directly use the Newton–Raphson method by using the built-in S-plus function `nonlinear minimization subject to box constraints` to speed up the convergence of the algorithm. We showed that the method provides an alternative for variable selection when the diseased and nondiseased groups are unbalanced in the dataset. In using this method for data analysis, a common cross-validation or independent validation is needed to be performed as usual.

ACKNOWLEDGEMENTS

This research was supported in part by ACS Institutional Research Grants IRG-97-153-07. The authors thank the editor and the referees for their constructive comments.

Received 11 December 2008

REFERENCES

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **96**, 6745–6750.

Böhning, D. and Lindsay, B. G. (1988). Monotonicity of quadratic approximation algorithms. *Annals of the Institute of Statistical Mathematics* **40**, 641–663.

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

De Leeuw, J. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis* **50**, 21–39.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. J. (2004). Least angle regression (with discussion). *The Annals of Statistics* **32**, 407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30**, 74–99.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**(2), 109–148.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**(3), 397–416.

Groenen, P. J. F., Giaquinto, P. and Kiers, H. L. (2003). Weighted majorization algorithms for weighted least squares decomposition models. Technical Report EI 2003–09, Econometrics Institute, Erasmus University, Rotterdam, Netherlands.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC, Boca Raton.

Hastie, T., Tibshirani, R. J. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Jaakkola, T. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**, 25–37.

Knight, K. and Fu, W. J. (2000). Asympotics for lasso-type estimators. *The Annals of Statistics* **28**, 1356–1378.

Lange, K., Hunter, D. R. and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics* **9**, 1–20.

Liu, C. H. (2000). Estimation of discrete distributions with a class of simplex constraints. *Journal of the American Statistical Association* **95**, 109–120.

Liu, Z. Q., Jiang, F., Tian, G. L., Wang, S., Sato, F., Meltzer, S. J. and Tan, M. (2007). Sparse logistic regression with L_p penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology* **6**(1), Article 6.

Madigan, D. and Ridgeway, G. (2004). Discussion on "Least angle regression" by Efron et al. *The Annals of Statistics* **32**, 465–469.

Malioutov, D. M., Cetin, M., and Willsky, A. S. (2005). A Sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Trans. Signal Processing* **53**(8), 3010–3022.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **37**, 362–372.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20**, 389–403.

Pepe, M. S. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press.

Stine, R. A. (2004). Discussion on "Least angle regression" by Efron et al. *The Annals of Statistics* **32**, 475–481.

Tan, M., Tian, G. L. and Fang, H. B. (2003). Estimating restricted normal means using the EM-type algorithms and IBF sampling. In *Development of Modern Statistics and Related Topics — In Celebration of Prof. Yaoting Zhang's 70th Birthday* (J. Huang and H. Zhang, Eds), 53–73. World Scientific, New Jersey.

Tan, M., Tian, G. L., Fang, H. B. and Ng, K. W. (2007). A fast EM algorithm for quadratic optimization subject to convex constraints. *Statistica Sinica* **17**(3), 945–964.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Wahba, G. (2007). Regularization methods in statistical model building: Statisticians, computer scientists, classification and machine learning. Presidential invited address, Joint Statistical Meetings, Salt Lake City, July 30, 2007.

Guo-Liang Tian
Department of Statistics and Actuarial Science,
The University of Hong Kong, Pokfulam Road,
Hong Kong, P.R. China
E-mail address: gltian@hku.hk

Hong-Bin Fang
Division of Biostatistics,
University of Maryland Greenebaum Cancer Center,
10 South Pine Street, MSTF Suite 261,
Baltimore, Maryland 21201, USA

Zhenqiu Liu
Division of Biostatistics,
University of Maryland Greenebaum Cancer Center,
10 South Pine Street, MSTF Suite 261,
Baltimore, Maryland 21201, USA

Ming T. Tan
Division of Biostatistics,
University of Maryland Greenebaum Cancer Center,
10 South Pine Street, MSTF Suite 261,
Baltimore, Maryland 21201, USA