



A New Algorithm in Maximum Likelihood Estimation for Generalized Linear Models

Yufang Wen (Corresponding author)
Department of Science, Yanshan University
Qinhuangdao 066004, China

Xiangdong Song
Department of Science, Yanshan University
Qinhuangdao 066004, China

Haisen Zhang
Department of Science, Yanshan University
Qinhuangdao 066004, China

Abstract

We introduce a new algorithm for L_1 regularized generalized linear models. The L_1 regularization procedure is useful, especially because it, in effect, selects variables according to the amount of penalization on the L_1 norm of the coefficients, in a manner less greedy than forward selection/backward deletion. The algorithm efficiently computes solutions along the entire regularization path using the predictor-corrector method of convex-optimization. Selecting the step length of the regularization parameter is critical in controlling the overall accuracy of the paths; we suggest intuitive and flexible strategies for choosing appropriate values.

Keywords: Generalized Linear Models, Predictor-Corrector algorithm

1. Introduction

In this paper we propose a predictor-corrector algorithm for L_1 regularized generalized linear models (GLM). GLM models a random variable Y that follows a distribution in the exponential family using a linear combination of the predictors, $x'\beta$, where x and β denote vectors of the predictors and the coefficients, respectively. The random and the systematic components may be linked through a non-linear function; therefore, we estimate the coefficient β by solving a set of non-linear equations that satisfy the maximum likelihood criterion.

$$\hat{\beta} = \arg \max_{\beta} L(y; \beta) \quad (1)$$

where L denotes the likelihood function with respect to the given data $\{(x_i, y_i) : i = 1, \dots, n\}$.

When the number of predictors p exceeds the number of observations n , or when insignificant predictors are present, we can impose a penalization on the L_1 norm of the coefficients for an automatic variable selection effect. Analogous to Lasso (Tibshirani 1996) that added a penalty term, to the squared error loss criterion, we modify criterion (1) with a regularization:

$$\hat{\beta}(\lambda) = \arg \min_{\beta} \{-\log L(y; \beta) + \lambda \|\beta\|_1\} \quad (2)$$

where $\lambda > 0$ is the regularization parameter. Logistic regression with L_1 penalization has been introduced and applied by other researchers, for example in Shevade & Keerthi (2003).

2. Problem setup

Let $\{(x_i, y_i) : x_i \in \mathcal{R}^p, y_i \in \mathcal{R}, i = 1, \dots, n\}$ be n pairs of p factors and a response. Y follows a distribution in the

exponential family with mean $\mu = E(Y)$ and variance $V = \text{Var}(Y)$. Depending on its distribution, the domain of y_i could be a subset of \mathfrak{R} . GLM models the random component Y by equating its mean μ with the systematic component η through a link function g :

$$\eta = g(\mu) = \beta_0 + x'\beta \tag{3}$$

The likelihood of Y is expressed as follows (McCullagh & Nelder 1989):

$$L(y; \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\} \tag{4}$$

$a(\cdot), b(\cdot)$, and $c(\cdot)$ are functions that vary according to the distributions. Assuming that the dispersion parameter ϕ is known, we are interested in finding the maximum likelihood solution for the natural parameter θ , and thus $(\beta_0, \beta)'$, with a penalization on the size of the L_1 norm of the coefficients ($\|\beta\|_1$). Therefore, our criterion with a fixed λ is reduced to finding $\beta = (\beta_0, \beta)'$, which minimizes the following:

$$l(\beta, \lambda) = -\sum_{i=1}^n \{y_i \theta(\beta)_i - b(\theta(\beta)_i)\} + \lambda \|\beta\|_1 \tag{5}$$

Assuming that none of the components of β is zero and differentiating $l(\beta, \lambda)$ with respect to β , we define a function H :

$$H(\beta, \lambda) = \frac{\partial l}{\partial \beta} = -X'W(y - \mu) \frac{\delta \eta}{\delta \mu} + \lambda \text{Sgn} \begin{pmatrix} 0 \\ \beta \end{pmatrix} \tag{6}$$

where X is an n by $p+1$ matrix including the column of 1's, W is a diagonal matrix with n diagonal elements $V_i^{-1}(\frac{\delta \mu}{\delta \eta})_i^2$, and $(y - \mu) \frac{\delta \eta}{\delta \mu}$ is a vector with n elements $(y_i - \mu_i)(\frac{\delta \eta}{\delta \mu})_i$.

Our goal is to compute the entire solution path for the coefficients β , with λ varying from ∞ to 0. We achieve this by drawing the uniquely determined curve $H(\beta, \lambda) = 0$ in $(p+2)$ dimensional space ($\beta \in \mathfrak{R}^{p+1}$ and $\lambda \in \mathfrak{R}_+$). Because $l(\beta, \lambda)$ is a convex function of β , there exists a $\beta(\lambda)$ that attains the unique minimum value for each $\lambda \in \mathfrak{R}_+$. In fact, a unique continuous and differentiable function $\beta(\lambda)$, such that $H(\beta(\lambda), \lambda) = 0$ exists within each open range of λ that yields a certain active set of variables; the existence of such mappings ($\lambda \rightarrow \beta(\lambda)$) can be shown using the implicit function theorem (Munkres 1991). We find the mapping $\beta(\lambda)$ sequentially with decreasing λ .

3. Predictor-Corrector algorithm

We introduced an algorithm that implements the predictor-corrector method to determine the entire path of the coefficient estimates as λ varies, i.e., to find $\{\hat{\beta}(\lambda); 0 < \lambda < \infty\}$. Starting from $\lambda = \infty$, our algorithm computes a series of solution sets, each time estimating the coefficients with a smaller λ based on the previous estimate. Each round of optimization consists of three steps: determining the step size in λ ; predicting the corresponding change in the coefficients, and correcting the error in the previous prediction.

The following lemma provides the initialization of the coefficient paths:

Lemma 1: *When λ exceeds a certain threshold, the intercept is the only nonzero coefficient: $\hat{\beta}_0 = g(\bar{y})$ and*

$$H((\hat{\beta}_0, 0 \dots, 0), \lambda) = 0 \text{ for } \lambda > \max_{j \in \{1, \dots, p\}} |x'_j \hat{W}(y - \bar{y}) g'(\bar{y})| \tag{7}$$

Proof. The Karush-Kuhn-Kuhn-Tucker (KKT) optimality conditions for minimizing (5) imply

$$\left| x'_j \hat{W}(y - \hat{\mu}) \frac{\delta \eta}{\delta \mu} \right| < \lambda \Rightarrow \hat{\beta}_j = 0 \text{ for } j = 1, \dots, p \tag{8}$$

When $\hat{\beta}_j = 0$ for all $j = 1, \dots, p$, the KKT conditions again imply

$$1' \hat{W}(y - \hat{\mu}) \frac{\delta \eta}{\delta \mu} = 0 \tag{9}$$

Which, in turn, yields $\hat{\mu} = \bar{y} = g^{-1}(\hat{\beta}_0)1$.

As λ is decreased further, other variables join the active set, beginning with the variable $j_0 = \arg \max_j |x'_j(y - \bar{y})|$.

Reducing λ , we alternate between a predictor and a corrector step.

3.1 Predictor step

In the k -th predictor step, $\beta(\lambda_{k+1})$ is approximated by

$$\hat{\beta}^{k+} = \hat{\beta}^k + (\lambda_{k+1} - \lambda_k) \frac{\partial \beta}{\partial \lambda} \tag{10}$$

$$= \hat{\beta}^{k+} - (\lambda_{k+1} - \lambda_k) \left(X_A' W_k X_A \right)^{-1} \text{Sgn}(0, \beta^k)' \tag{11}$$

W_k and X_A denote the current weight matrix and the columns of X for the factors in the current active set, respectively. β in the above equations are composed only of current nonzero coefficients. This linearization is equivalent to making a quadratic approximation of the log-likelihood and extending the current solution $\hat{\beta}^k$ by taking a weighted lasso step (as in LARS).

Define $f(\lambda) = H(\beta(\lambda), \lambda)$; in the domain that yields the current active set, $f(\lambda)$ is zero for all λ . By differentiating f with respect to λ , we obtain

$$f'(\lambda) = \frac{\partial H}{\partial \lambda} + \frac{\partial H}{\partial \beta} \frac{\partial \beta}{\partial \lambda} = 0 \tag{12}$$

from which we compute $\delta\beta/\delta\lambda$.

3.2 Corrector step

In the following corrector step, we use $\hat{\beta}^{k+}$ as the initial value to find the β that minimizes $l(\beta, \lambda_{k+1})$, as defined in (5) (i.e., that solves $H(\beta, \lambda_{k+1}) = 0$ for β). Any (convex) optimization method that applies to the minimization of a differentiable objective function with linear constraints may be implemented. The previous predictor step has provided a warm start; because $\hat{\beta}^{k+}$ is usually close to the exact solution $\hat{\beta}^{k+1}$, the cost of solving for the exact solution is low. The corrector steps not only find the exact solution at a given λ but also yield the directions of β for the subsequent predictor steps.

3.3 Active set

The active set A begins from the intercept as in Lemma 3.1; after each corrector step, we check to see if A should have been augmented. The following procedure for checking is justified and used by Rosset & Zhu (2003) and Rosset (2004):

$$\left| x_j' W (y - \mu) \frac{\delta\eta}{\delta\mu} \right| > \lambda \text{ for any } j \in A \Rightarrow A \leftarrow A \cup \{j\} \tag{13}$$

We repeat the corrector step with the modified active set until the active set is not augmented further. We then remove the variables with zero coefficients from the active set. This is,

$$\left| \hat{\beta}_j \right| = 0 \text{ for any } j \in A \Rightarrow A \leftarrow A - \{j\} \tag{14}$$

3.4 Step length

Two natural choices for the step length $\Delta_k = \lambda_k - \lambda_{k+1}$ are:

- (1) $\Delta_k = \Delta$, fixed for every k , or
- (2) a fixed change L in L_1 arc-length, achieved by setting $\Delta_k = L / \|\delta\beta/\delta\lambda\|_1$.

As we decrease the step size, the exact solutions are computed on a finer grid of λ values, and the coefficient path becomes more accurate.

We propose a more efficient and useful strategy:

- (3) select the smallest Δ_k that will change the active set of variables.

We give an intuitive explanation of how we achieve this, by drawing on analogies with the LARS algorithm (Efron et al. 2004). At the end of the k -th iteration, the corrector step can be characterized as finding a weighted Lasso solution that satisfies $-X_A' W_k (y - \mu) \frac{\delta\eta}{\delta\mu} + \lambda_k \text{Sgn}\left(\begin{matrix} 0 \\ \beta \end{matrix}\right) = 0$. This weighted Lasso also produces the direction for the next

predictor step. If the weights W_k were fixed, the weight Lars algorithm would be able to compute the exact step length to the next active-set change point. We use this step length, even though in practice the weights change as the path progresses.

Lemma 2: Let $\hat{\mu}$ be the estimates of y from a corrector step, and denote the corresponding weighted correlations as

$$\hat{c} = X' \hat{W} (y - \hat{\mu}) \frac{\delta \eta}{\delta \mu}. \tag{15}$$

The absolute correlations of the factors in A (except for the intercept) are λ , while the values are smaller than λ for the factors in A^c .

Proof. The Karush-Kuhn-Tucker (KKT) optimality for minimizing (5) imply

$$\hat{\beta}_j \neq 0 \Rightarrow \left| x'_j \hat{W} (y - \hat{\mu}) \frac{\delta \eta}{\delta \mu} \right| = \lambda. \tag{16}$$

This condition, combined with (7) and (8), proves the argument.

The next predictor step extends $\hat{\beta}$ as in (11), and, thus, the current correlations change. Denoting the vector of changes in correlation for a unit decrease in λ as a ,

$$c(h) = \hat{c} - ha \tag{17}$$

$$= \hat{c} - h X' \hat{W} X_A (X'_A \hat{W} X_A)^{-1} \text{Sgn} \begin{pmatrix} 0 \\ \hat{\beta} \end{pmatrix}, \tag{18}$$

Where $h > 0$ is a given decrease in λ . For the factors in A , the values of a are those of $\text{Sgn} \begin{pmatrix} 0 \\ \hat{\beta} \end{pmatrix}$. To find the h

with which any factor in A^c yields the same absolute correlation as the ones in A , we solve the following equations:

$$|c_j(h)| = |\hat{c}_j - ha_j| = \lambda - h \text{ for any } j \in A^c \tag{19}$$

The equations suggest an estimate of the step length in λ as

$$h = \min_{j \in A^c} \left\{ \frac{\lambda - \hat{c}_j}{1 - a_j}, \frac{\lambda + \hat{c}_j}{1 + a_j} \right\}. \tag{20}$$

In addition, to check if any variable in the active set reaches 0 before λ decreases by h , we solve the equations

$$\beta_j(\tilde{h}) = \hat{\beta}_j + \tilde{h} (X'_A \hat{W} X_A)^{-1} \text{Sgn} \begin{pmatrix} 0 \\ \hat{\beta} \end{pmatrix} = 0 \text{ for any } j \in A. \tag{21}$$

If $0 < \tilde{h} < h$ for any $j \in A$, we expect that the corresponding variable will be eliminated from the active set before any other variable joins it; therefore, \tilde{h} rather than h is used as the next step length.

Letting the coefficient paths be piecewise linear with the knots placed where the active set changes is a reasonable simplification of the truth based on our experience (using both simulated and real datasets). If the smallest step length that modifies the active set were to be larger than the value we have estimated, the active set remains the same, even after the corrector step. If the true step length were smaller than expected, and, thus, we missed the entering point of a new active variable by far, we would repeat a corrector step with an increased λ . Therefore, our path algorithm almost precisely detects the values of λ at which the active set changes, in the sense that we compute the exact coefficients at least once before their absolute values grow larger than δ (a small fixed quantity).

We can easily show that in the case of Gaussian distribution with the identity link, the piecewise linear paths are exact. Because $\hat{\mu} = X' \hat{\beta}$ and $V_i = \text{Var}(y_i)$ is constant for $i = 1, \dots, n$, $H(\beta, \lambda)$ implies to $-X'(y - \mu) + \lambda \text{Sgn}(0, \beta)'$. The step lengths are computed with no error; in addition, since the predictor steps yield the exact coefficient values, corrector steps are not necessary. In fact, the paths are identical to those Lasso.

4. Discussion

We can extend the use of the predictor-corrector scheme by generalizing the loss+penalty function to any convex and almost differentiable functions. For example, we can find the entire regularization path for the Cox proportional hazards models with L_1 penalization. Just as the solution paths for Gaussian distribution were computed with no error through the predictor-corrector method, so any other piecewise linear solution paths can be computed exactly by

applying the same strategy.

References

R.Tibshirani (1997), The lasso method for variable selection in the cox models, *Statistics in Medicine*, pp: 385-395.

Osborne, M., Presnell, B. & Turlach, B. (2000), On the lasso and its dual, *Journal of Computational and Graphical Statistics* pp. 319-337.

Zou, H & Hastie, T. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B*, pp:301-320.

M.Osborne,B. Presnell & B.Turlach. (2002), A new approach to variable selection in least squares problems, *IMA Journal of Numerical Analysis*, 20(3): 389-403.

Xiru Chen (2003), Generalized Linear Models(5), *Mathematical Statistics and Management*, 22(3), pp:56-63.

S.Shevade & S.Keerthi (2003). *A simple and efficient algorithm for gene selection using sparse logistic regression*, *Bioinformatics*, pp:2246-2253.