Review: Multiplicative iterative path algorithm for non-negative generalized linear model

Jieun Shin

May 24, 2023

1 Background

1.1 generalized linear model

For generalized linear models (GLM), a random variable Y follows a distribution belonging to the exponential family and its mean value μ is related to the linear predictor $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$, where $\boldsymbol{x}$ is a vector for the data of a set of explanatory variables and $\boldsymbol{\beta}$ denotes a vector for the regression coefficients. Usually, one assumes that the mean and the linear predictor are linked through a link function g, i.e., $g(\mu) = \eta$. The regression coefficients are estimated by maximizing the log-likelihood:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \log L(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}), \tag{1}$$

where $\log L$ denotes the log-likelihood derived from the data vector $\boldsymbol{y}$ and $\boldsymbol{X}$ is a matrix combining all the $\boldsymbol{x}$'s.

1.2 LASSO

LASSO is the regularized regression for simultaneously performing variable selection and shrinkage for least-squares regressions when some important variables are to be selected out of p predictor variables. LASSO can be conceived as adopting a penalty using the l_1-norm of the regression coefficients, where the l_1-norm penalty can efficiently enforce sparsity. This concept was extended to GLMs by adding an l_1 penalty to the log-likelihood function. Without loss of generality, we assume that the predictor variables are standardized i.e., $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = 1$ for $j = 1, 2, \ldots, p$. Then, l_1 regularized estimates in GLM are given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \log L(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) - \lambda ||\boldsymbol{\beta}||_1, \tag{2}$$

where $\lambda > 0$ is the regularization parameter. Park and Hestie (2007) proposed an algorithm to compute the entire regularization path of (2) using a predictor corrector method. Friedman et al. (2010) developed a highly efficient coordinate descent method to solve (2).

회귀계수를 non-negative하게 추정해야 하는 문제는 종종 제기되어 왔다 (non-negative least squares 등)

2 Multiplicative iterative algorithm for penalized non-negative GLMs

2.1 l_1 penalized non-negative GLMs

The constrained optimization of our interest is

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \geq 0}{\operatorname{argmax}} \ell(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) - \lambda ||\boldsymbol{\beta}||_1, \tag{3}$$

where $l = \log L$ and the constraint $\boldsymbol{\beta} \geq 0$ is interpreted elementwise. Since $\boldsymbol{\beta} \geq 0$, the objective function in (3) becomes

$$\phi(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) - \lambda \boldsymbol{\beta}^T \mathbf{1}, \tag{4}$$

where $\mathbf{1}$ denotes a vector of ones with the same size as $\boldsymbol{\beta}$. The KKT condition for the constrained optimization problem (3) are

$$\frac{\partial \phi}{\partial \beta_j} = 0 \text{ if } \beta_j > 0 \text{ and } \frac{\partial \phi}{\partial \beta_j} < 0 \text{ if } \beta_j = 0. \tag{5}$$

Therefore, we wish to solve the following simultaneous equations

$$\beta_j \left(\frac{\partial \ell}{\partial \beta_j} - \lambda \right) = 0, \tag{6}$$

where $j = 1, 2, \ldots, p$ and subject to all $\beta_j \geq 0$. In (6) and for the GLM model considered by this paper, $\partial \ell / \partial \beta_j = \boldsymbol{x}^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{\mu})$. Writing $\partial \ell / \partial \beta_j = (\partial \ell / \partial \beta_j)^+ + (\partial \ell / \partial \beta_j)^-$, where we can devise the following iteration to update $\beta_j, j = 1, 2, \ldots, p$ (development of Ma (2010)):

$$\beta_j^{(k+\frac{1}{2})} = \beta_j^{(k)} \frac{\left(\frac{\partial \ell(\boldsymbol{\beta}^{(k)})}{\partial \beta_j} \right)^+ + \epsilon}{\lambda - \left(\frac{\partial \ell(\boldsymbol{\beta}^{(k)})}{\partial \beta_j} \right)^- + \epsilon}, \tag{7}$$

where function $\partial \ell(\boldsymbol{\beta}^{(k)}) / \partial \beta_j$ denotes $\partial \ell(\boldsymbol{\beta}) / \partial \beta_j$ with $\boldsymbol{\beta}$ replaced by its current estimate $\boldsymbol{\beta}^{(k)}$ and $\epsilon > 0$ is a small threshold used to avoid zero denominator. Note that both numerator and denominator in (7) are nonnegative and hence $\beta_j^{(k+\frac{1}{2})} \geq 0$ provided $\beta_j^{(k)} \geq 0$. The MI algorithm given by (7) may not converge as the objective function $\phi(\boldsymbol{\beta})$ may fail to increase when $\boldsymbol{\beta}$ moves from $\boldsymbol{\beta}^{(k)}$ to $\boldsymbol{\beta}^{(k+\frac{1}{2})}$. To overcome this problem, a line search step must be included. We first rewrite (7) in the form of a gradient algorithm:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + s_j^{(k)} \frac{\partial \phi(\boldsymbol{\beta}^{(k)})}{\partial \beta_j}, \tag{8}$$

where $s_j^{(k)} = \beta_j^{(k)} / (\lambda - [\partial \phi(\boldsymbol{\beta}^{(k)}) / \partial \beta_j]^- + \epsilon)$. Then a line search step, such as the Armijo line search strategy, can be introduced to ensure that $\phi(\boldsymbol{\beta})$ is not reduced at each iteration. After incorporating a line search, β_j is updated according to

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + w^{(k)} s_j^{(k)} \frac{\partial \phi(\boldsymbol{\beta}^{(k)})}{\partial \beta_j}, \tag{9}$$

where $0 < w^{(k)} \leq 1$ is the step size determined by the line search. To be specific, the Armijo line search is a finite terminating algorithm which starts with $w = 1$, and for each w it checks if the following Armijo condition is satisfied:

$$\phi(\boldsymbol{\beta}^{(k)} + w \boldsymbol{d}^{(k)}) \geq \phi(\boldsymbol{\beta}^{(k)}) + \xi w \left(\frac{\partial \phi(\boldsymbol{\beta}^{(k)}}{\partial \boldsymbol{\beta}} \right)^T \tag{10}$$

where $\boldsymbol{d}^{(k)}$ represents a p-vector with elements $d_j^{(k)} = s_j^{(k)} \partial \phi(\boldsymbol{\beta}^{(k)}) / \partial \beta_j$ and $0 < \xi < 1$ is a fixed parameter such as $\xi = 10^{-2}$.

2.2 Elastic net GLM with non-negative coefficients

When some of the predictor variables are highly correlated, regression coefficient estimates can be very unstable. Under such a situation, the elastic net penalty for linear regression model can be very useful. Elastic net imposes penalties using the l_2-norm of the regression coefficients in addition to the penalty on their $l_!$-norm.

The same idea was extended to GLM by Park and Hestie (2007) in the case of correlated predictors. Adopting the same idea, a version of elastic net for GLMs with non-negative coefficients is proposed here when predictors are strongly correlated, namely the non-negatively constrained regression coefficients are estimated by maximizing the following objective function

$$\phi(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) - \lambda_1 \boldsymbol{\beta}^T \mathbf{1} - \frac{\lambda_2}{2} \boldsymbol{\beta}^T \boldsymbol{\beta}, \tag{11}$$

subject to $\boldsymbol{\beta} \geq \mathbf{0}$, where $\lambda_1 > 0$ and $\lambda_2 > 0$ and the regularization parameters. Parameter λ_2 is generally kept small and fixed ensuring that the stability of the fit is not affected by strong correlations between predictors. When the predictor correlations are not strong, the quadratic penalty component has negligible effect because of a small λ_2. (11) can also be written as

$$\phi(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) - \lambda[\tau \boldsymbol{\beta}^T \mathbf{1} - (1-\tau)\boldsymbol{\beta}^T \boldsymbol{\beta}], \tag{12}$$

where $0 \geq \tau \geq 1$. Clearly, $\tau = 1$ corresponds to the lasso penalty.

It is easy to check that the MIA updating scheme for the case of elastic net penalty is now given by

$$\beta_j^{(k+1)} = \beta_j^{(k)} + w^{(k)} s_j^{(k)} \left(\frac{\partial \ell(\boldsymbol{\beta}^{(k)})}{\partial \beta_j} - \lambda_1 - \lambda_2 \beta_j^{(k)} \right), \tag{13}$$

where $s_j^{(k)} = \beta_j^{(k)} / \{[\ell(\boldsymbol{\beta}^{(k)})/\beta_j]^- + \lambda_1 + \lambda_2 \beta_j^{(k)} + \epsilon\}$ with ϵ being a small constant such as 10^{-2}. Line search step size $w^{(k)}$ can be again determined by the Armijo method.

3 reference

[1] Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(4), 659-677.

[2] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1), 1.

[3] Ma, J. (2010). Positively constrained multiplicative iterative algorithm for maximum penalized likelihood tomographic reconstruction. IEEE Transactions on Nuclear Science, 57(1), 181-192.