Proximal Methods for Lasso Penalties in the Cox Proportional Hazards Model

José L. Jiménez[1], *José R. Dorronsoro*[2]

[1]joseluis.jimenezm@estudiante.uam.es,
Departmento de Ingeniería Informática, Universidad Autónoma de Madrid
[2]jose.dorronsoro@uam.es,
Departmento de Ingeniería Informática, Universidad Autónoma de Madrid

Abstract

With the emergence of new biomedical technologies, statistical methods for the analysis of high dimensional survival data have become increasingly important. In this work we aim to provide a general overview of the application of regularization methods to the Cox PHM for automatic variable selection, and show how this regularization may be done under a proximal optimization paradigm, which to our knowledge, is a novel approximation to the problem.

Keywords: cox proportional hazards model; lasso; FISTA.

AMS: 62N99, 68T99.

1. Introduction

In the last few years, new technologies in the field of genomics have led to a great amount of biomedical data. Gene expression data have changed our understanding of complex diseases such as cancer. However, it has a main characteristic: the number of features greatly exceeds the number of observations. As a result, many classical statistical approaches cannot be applied to these data without major modifications.

Moreover, when there is survival data avilable, we may be interested in assessing which features are most associated with a survival outcome.

Consider a traditional survival analysis framework with data of the form $(y_1, x_1, \delta_1), ..., (y_n, x_n, \delta_n)$, where y_i is the observed time of failure if $\delta_i = 1$, or is right-censored if $\delta_i = 0$. The vector $\mathbf{x}_i$ contains the features $(x_{i1}, x_{i2}, ..., x_{ip})$. Assuming no ties and letting $t_1 < t_2 < ... < t_m$ be the increasing list of unique failure times, the Cox Proportional Hazards Model (PHM) assumes a semi-parametric form for the hazard defined as

$$h(t|\mathbf{w}) = h_0(t) \exp(\mathbf{w}^T \mathbf{x}), \tag{1}$$

where $h_0(t)$ is a completely unspecified baseline hazard function, and $\mathbf{w}^T = (w_1, w_2, ..., w_p)^T$ is an unknown vector of regression coefficients. Note that in (1) we have decomposed the hazard into a product of two elements, where $h_0(t)$, depends on time but not on the covariates, and $\exp(\mathbf{w}^T \mathbf{x})$, which depends on the covariates but not on time.

We want to fit (1) to a given sample and estimate the optimal $\mathbf{w}$ parameters. For the estimation of these regression parameters, we propose to minimize the minus partial log likelihood function, a convex defined as

$$l(\mathbf{w}) = -\sum_{i=1}^{N}\left[\mathbf{w}^T\mathbf{x}_i - \log\left(\sum_{j\epsilon R(t_i)}\exp\left(\mathbf{w}^T\mathbf{x}_j\right)\right)\right]. \tag{2}$$

This minimization problem is equivalent to maximizing the partial log likelihood function proposed by [1]. Because (2) is convex, its solution is unique and the minimization may be done using various algorithms, where perhaps the most common is the Newton Raphson algorithm, a standard tool for solving unconstrained smooth optimization problems.

2. Algorithm

Considering the framework described in section 1, we are interested in a regularized version of the Cox PHM [6], where a Lasso penalty is added to the minus log partial likelihood function for shrinkage and variable selection. This method is based on the Lasso [5], which was originally designed for the linear regression problem. In this work, we have the following unconstrained minimization problem.

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\left\{-\sum_{i=1}^{N}\left[\mathbf{w}^T\mathbf{x}_i - \log\left(\sum_{j\epsilon R(t_i)}\exp\left(\mathbf{w}^T\mathbf{x}_j\right)\right)\right] + \lambda\left\|\mathbf{w}\right\|_1\right\}, \tag{3}$$

where λ controls the amount of shrinkage applied to the coefficients.

Several algorithms have been proposed for solving the reguarized Cox PHM. [6] proposes to solve (3) through constrained reweighted least squares. This approach is very similar to the traditional Newton Raphson updates, although the Lasso constrain is added to the problem. [2] approaches the problem with an algorithm based on a combination of gradient ascent optimization with the traditional Newton Raphson updates. A characteristic of this approach is that it follows the gradient of the likelihood from a given starting point using the full gradient at each step. [2] has available an algorithms named `penalized`, freely available at CRAN.

In this work we propose solving (3) using FISTA, an algorithm that uses the proximal operator. Proximal algorithms may be viewed as an analogous tool for non-smooth, large scale optimization problems [4]. They are very applicable and well-suited to problems of recent interest involving high-dimensional datasets. Even though the most popular application of proximal algorithms is for solving the Lasso, with a minor modification we can make it solve the regularized Cox PHM given that it only requires to know the function we want to minimize, and its gradient. Our minimization problem (3) is hence formed by two convex functions, $f(\mathbf{w}) = l(\mathbf{w})$ and $g(\mathbf{w}) = \left\|\mathbf{w}\right\|_1$, where $f(\mathbf{w})$ is differentiable and Lipschitz continuous and $g(\mathbf{w})$ is non-differentiable.

Following [4], a definition for the proximal operator is given by

$$\text{prox}_g(\mathbf{w}_k) = \underset{\mathbf{w}_{k+1}}{\arg\min} \left\{ \frac{1}{2} \|\mathbf{w}_{k+1} - \mathbf{w}_k\|_2^2 + \|\mathbf{w}_{k+1}\|_1 \right\}. \tag{4}$$

Equation (4) gives already an idea of the iterative procedure we are going to employ. Consider $\mathbf{w}_k$ as some possible values for $g(\mathbf{w})$. Of course, they are not the minimum, but they can be though as a starting point. Now we need to find some $\mathbf{w}_{k+1}$ values that are a middle point between $\mathbf{w}_k$ and the minimum of $g(\mathbf{w})$. In other word, we take one step towards the minimum of the $g(\mathbf{w})$. If we repeat this procedure, it is possible to show that it will converge to the solution of the problem.

However, (4) only minimizes the $g(\mathbf{w})$. Our interest lies in minimizing (3) which is composed by the sum of two convex functions. For this task, we can re-write (4) as

$$\text{prox}_{\frac{\gamma}{L}g}\left(\mathbf{w}_k - \frac{\gamma}{L}\nabla f(\mathbf{w}_k)\right) = \underset{\mathbf{w}_{k+1}}{\arg\min} \left\{ \frac{1}{2} \left\|\mathbf{w}_{k+1} - \left(\mathbf{w}_k - \frac{\gamma}{L}\nabla f(\mathbf{w}_k)\right)\right\|_2^2 + \|\mathbf{w}_{k+1}\|_1 \right\}, \tag{5}$$

where L is the Lipschitz constant of $\nabla f(\mathbf{w})$, and γ controls the step size.

The interpretation of (5) is similar to the interpretation of (4). The proximal operator again minimizes $g(\mathbf{w})$. However, notice that at each step, the algorithm receives $\mathbf{w}_k - \frac{\gamma}{L}\nabla f(\mathbf{w}_k)$, rather than just $\mathbf{w}_k$. It is easy to see that $\mathbf{w}_k - \frac{\gamma}{L}\nabla f(\mathbf{w}_k)$ is a traditional gradient descent step, which means that we first do a gradient descent step, controlled by γ, towards the minimum of $f(\mathbf{w})$, and then, with the result given by the gradient descent step, we take a step towards the minimum of $g(\mathbf{w})$. This way, we sequentially minimize both functions at the same time.

This algorithm is referred as "Iterative Shrinkage Thresholding Algorithm" or ISTA, and includes a backtracking step if the Lipschitz constant is unknown and needs to be approximated. An interesting property is that ISTA has a convergence rate equal to $O(1/k)$, which is similar to gradient descent. However, we can improve its convergence rate to $O(1/k^2)$ by mean of the Nesterov's accelerated gradient [3]. ISTA with the Nesterov's accelerated gradient receives the name of FISTA, which stands for "Fast Iterative Shrinkage Thresholding Algorithm".

3. Experiments

To illustrate the performance of our proposal, we test it using a genomic dataset with the competing algorithm in the field of regularized Cox PHM: `penalized`. This algorithm has been already introduced in section 2.

We are interested in assessing the number of iterations and time until convergence of each algorithm. Results are showed at Fig. 1 and we can see that FISTA outperforms `penalized` in both time and iterations until convergence. Notice that, because the convex nature of the problem, there is a unique solution for a given λ value, and hence an accuracy comparison is not needed.

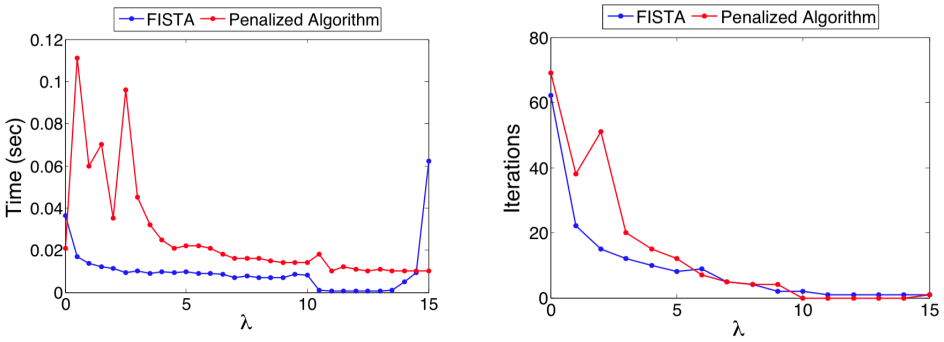

Figure 1: Time and iterations until convergence of FISTA (blue) and `penalized` algorithm (red) for different values of λ.

4. Conclusions

In this work we propose to solve the regularized Cox PHM under a proximal optimization paradigm, which to our knowlege, is a novel approach to this problem. Proximal methods were originally developed for the Lasso, but can be easily modified for our purpose. We show that proximal methods offer not only a natural solution of the problem, which is unique given its convex nature, but require less time and iterations to converge than a competing algorithm in this field.

5. Bibliography

[1] David, Cox R. "Regression models and life tables." Journal of the Royal Statistical Society 34 (1972): 187-220.

[2] Goeman, J. J. (2010). *L1 penalized estimation in the cox proportional hazards model*. Biometrical Journal, 52(1), 70-84.

[3] Nesterov, Y. (1983, February). *A method of solving a convex programming problem with convergence rate O (1/k2)*. In Soviet Mathematics Doklady (Vol. 27, No. 2, pp. 372-376).

[4] Parikh, N., & Boyd, S. (2013). *Proximal algorithms*. Foundations and Trends in optimization, 1(3), 123-231.

[5] Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society*. Series B (Methodological), 267-288.

[6] Tibshirani, R. (1997). *The lasso method for variable selection in the Cox model*. Statistics in medicine, 16(4), 385-395.