Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. doi: 10.18637/jss.v000.i00

Regularization Paths for Interval Regression Models via Coordinate Descent

Anuj Khare
IIT Guwahati

Toby Dylan Hocking
McGill University

Jelle Goeman
Leiden University

Abstract

We describe a coordinate descent algorithm for estimating interval regression models with elastic net penalties. The output data we consider are of four possible censoring types: left, right, interval, and none. The models we consider are Gaussian, Logistic, TODO. We evaluate the speed and accuracy of the coordinate descent method on several real data sets, including survival prediction and and penalty learning problems.

Keywords: keywords, comma-separated, not capitalized, Java.

1. Introduction

There has been a recent interest in Accelerated Failure Time models (Huang, Ma, and Xie 2005; Cai, Huang, and Tian 2009; Khan and Shaw 2013).

The coordinate descent algorithm has been used for elastic net regularized generalized linear models (Friedman, Hastie, and Tibshirani 2010), including Cox's proportional hazards model (Simon, Friedman, Hastie, and Tibshirani 2011).

2. Model

In interval regression problems we are given n observations. For each observation $i \in \{1, \dots, n\}$, we have a p-dimensional input feature vector $\mathbf{x}_i \in \mathbb{R}^p$. The output data are lower $\underline{y}_i \in \{-\infty, \mathbb{R}\}$ and upper $\overline{y}_i \in \{\mathbb{R}, \infty\}$ limits which verify $\underline{y}_i \leq \overline{y}_i$.

For example, usual linear regression is a special case where $\underline{y}_i = \overline{y}_i \in \mathbb{R}$ for all observations $i \in \{1, \dots, n\}$. In this case we say that none of the observations are censored.

Another example is survival data, where each observation i corresponds to a patient who is administered a treatment in a study. The model predicts the amount of time that a patient survives after treatment, and there are two types of observed output data:

- Patient i died before the end of the study, at time $t_i > 0$ after treatment. This implies a log survival time of $\underline{y}_i = \overline{y}_i = \log t_i$. There is no censoring for this observation.
- Patient i was alive at the end of the study, and the time between treatment and the end of the study is t_i. This implies the patient's log survival time is somewhere in the interval $(\underline{y}_i, \overline{y}_i) = (\log t_i, \infty)$. This observation is right censored.

A final example is penalty learning, where the model predicts a positive penalty constant that is used to select the number of change-points in segmentation models (Rigaill, Hocking, Vert, and Bach 2013). The output data are intervals $(\underline{y}_i, \overline{y}_i)$ which are either finite or infinite, but for these data we never have $\underline{y}_i = \overline{y}_i$. In terms of the survival literature, we say that all of the observations are either left, right, or interval censored.

2.1. The concave maximum likelihood problem

The predicted output is $\beta + \mathbf{w}^\intercal \mathbf{x}_i$, where $\beta \in \mathbb{R}$ is the bias/intercept, and $w \in \mathbb{R}^p$ is a weight vector.

TODO discuss likelihood in terms of the scale $\sigma > 0$ and log-scale $s = \log \sigma \in \mathbb{R}$.

2.2. The convex minimization problem

The regularized model can be written as the following minimization problem.

$$\operatorname*{minimize}_{\mathbf{w},\beta,s} \text{Cost}_{\alpha,\lambda}(\mathbf{w}, \beta, s). \tag{1}$$

The Cost function has two regularization hyper-parameters, $\lambda \geq 0$ and $\alpha \in [0, 1]$, which must be fixed before solving for the optimization variables $\mathbf{w}, \beta, s$. The Cost function is composed of a smooth convex loss $\mathcal{L}$ and a convex penalty P_α.

$$\text{Cost}_{\alpha,\lambda}(\mathbf{w}, \beta, s) = \mathcal{L}(\mathbf{w}, \beta, s) + \lambda P_\alpha(\mathbf{w}). \tag{2}$$

The penalty $P_\alpha(\mathbf{w}) = \alpha||\mathbf{w}||_1 + (1 - \alpha)||\mathbf{w}||_2^2/2$ is the elastic net, and the average loss $\mathcal{L}$ is defined as

$$\mathcal{L}(\mathbf{w}, \beta, s) = \frac{1}{n} \sum_{i=1}^{n} \ell(\underline{y}_i, \overline{y}_i, s, \beta + \mathbf{w}^\intercal \mathbf{x}_i). \tag{3}$$

Let $\eta_i = \beta + \mathbf{w}^\intercal \mathbf{x}_i$. The observation-specific loss ℓ is defined as

$$\ell(\underline{y}, \overline{y}, s, \eta) = \begin{cases} s - \log f\left(\frac{y-\eta}{e^s}\right) & \text{if } \underline{y} = \overline{y} \\ -\log\left[F\left(\frac{\overline{y}-\eta}{e^s}\right) - F\left(\frac{\underline{y}-\eta}{e^s}\right)\right] & \text{otherwise.} \end{cases} \tag{4}$$

The losses for left and right censored observations can be obtained by setting $\underline{y} = -\textbf{Inf}$ and $\overline{y} = \textbf{Inf}$ respectively. The density f and cumulative distribution function F depend on the choice of the distribution (normal, logistic, etc). TODO table with precise equations for f and F for the models we have implemented.

3. Algorithm

We adapt the gradient computations of Therneau (2012). TODO

4. Results

We use the code to learn penalty functions that predict the number of change-points in segmentation models (Rigaill *et al.* 2013; Hocking, Rigaill, and Bourque 2015). These problems are relatively small in terms of number of observations n and features p.

TODO: is there a really big (in terms of observations or features) survival data set we could apply our code to?

References

Cai T, Huang J, Tian L (2009). "Regularized Estimation for the Accelerated Failure Time Model." *Biometrics*, **65**(2), 394–404.

Friedman JH, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*.

Hocking TD, Rigaill G, Bourque G (2015). "PeakSeg: constrained optimal segmentation and supervised penalty learning for peak detection in count data." In *Proc. 32nd ICML*, pp. 324–332.

Huang J, Ma S, Xie H (2005). "Regularized Estimation in the Accelerated Failure Time Model with High Dimensional Covariates." *Technical report*, University of Iowa, Department of Statistics and Actuarial Science.

Khan MHR, Shaw JEH (2013). "Variable Selection for Survival Data with A Class of Adaptive Elastic Net Techniques." ArXiv:1312.2079.

Rigaill G, Hocking T, Vert JP, Bach F (2013). "Learning sparse penalties for change-point detection using max margin interval regression." In *Proc. 30th ICML*, pp. 172–180.

Simon N, Friedman JH, Hastie T, Tibshirani R (2011). "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software*.

Therneau TM (2012). "A Package for Survival Analysis in S." Sections 6.7–6.9.

Affiliation:

Achim Zeileis
Department of Statistics and Mathematics
Faculty of Economics and Statistics
Universität Innsbruck
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@uibk.ac.at
URL: http://eeecon.uibk.ac.at/~zeileis/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd