Statistics and Probability Letters 109 (2016) 189–193

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Data-driven ridge regression for Aalen's additive risk model

Audrey Boruvka [a], Glen Takahara [a], Dongsheng Tu [a,b,*]

[a] Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, K7L 3N6, Canada
[b] NCIC Clinical Trials Group, Queen's University, 10 Stuart Street, Kingston, Ontario, K7L 3N6, Canada

ARTICLE INFO

Article history:
Received 12 December 2014
Received in revised form 11 November 2015
Accepted 11 November 2015
Available online 30 November 2015

Keywords:
Additive risk model
Event history data
L-curve
Mean square error
Ridge regression
Survival analysis

ABSTRACT

Two data-driven procedures, based respectively on the L-curve and generalized cross-validation, are proposed for ridge regression under Aalen's additive risk model. Monte Carlo simulations show that the L-curve is a useful criterion for identifying a nominal degree of regularization that appreciably reduces variance, particularly in smaller samples.

1. Introduction

Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ be the intensity of a multivariate process $\mathbf{N} = (N_1, \ldots, N_n)$ counting the occurrence of some event among n individuals. Aalen's (1978, 1980) additive risk model takes the form

$$\boldsymbol{\lambda}(t) = \mathbf{Y}(t)\boldsymbol{\alpha}(t), \quad 0 \le t \le \tau < \infty, \tag{1.1}$$

where $\mathbf{Y} = (\mathbf{Y}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_p)$ is a bounded, predictable process that essentially amounts to a time-varying design matrix of ones $\mathbf{Y}_0(t)$ and p ($n > p + 1$) covariates $\mathbf{Y}_1(t), \ldots, \mathbf{Y}_p(t)$, with entries for any individuals not at risk at time t- set to zero. Here the estimand is $\mathbf{A} = \int \boldsymbol{\alpha}$, a $(p + 1)$-vector of cumulative regression functions. These are nonparametric in the sense that $\int \mathbf{Y} \mathrm{d}\,\mathbf{A}$ should be non-decreasing, but otherwise the dependence of $\mathbf{A}$ on time is left unspecified. Aalen (1980) proposed an unconstrained ordinary least squares (OLS) estimator

$$\hat{\mathbf{A}}(t) = \int_0^t \mathbf{Y}(s)^- \mathrm{d}\,\mathbf{N}(s),$$

where

$$\mathbf{Y}(t)^- = J(t)\left(\mathbf{Y}(t)^\top \mathbf{Y}(t)\right)^{-1} \mathbf{Y}(t)^\top$$

is the Moore–Penrose inverse multiplied by the indicator $J(t)$ that this pseudo-inverse exists. OLS becomes numerically unstable whenever $\mathbf{Y}(t)^\top \mathbf{Y}(t)$ is near-singular. To increase stability and dispense with J, Aalen et al. (2004) considered a

* Corresponding author at: Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, K7L 3N6, Canada.
E-mail address: dtu@ctg.queensu.ca (D. Tu).

http://dx.doi.org/10.1016/j.spl.2015.11.010
0167-7152/

ridge regression (Hoerl and Kennard, 1970) estimator $\hat{\mathbf{A}}_{\text{RR}}$, obtained by replacing $\mathbf{Y}(t)^-$ with the regularized inverse

$$\mathbf{Y}(t)^-_{\text{RR}} = \left(\mathbf{Y}(t)^\top \mathbf{Y}(t) + \eta(t)\mathbf{I}\right)^{-1} \mathbf{Y}(t)^\top, \tag{1.2}$$

where $\eta(t)$ is a scalar regularization parameter and $\mathbf{I}$ is a $(p+1)\times(p+1)$ identity matrix. Generally the value of $\eta(t)$ should be large enough to ensure that $\mathbf{Y}(t)^-_{\text{RR}}$ is defined, but sufficiently small to avoid introducing much bias. Ridge parameter selection has been previously examined for the additive risk model in which covariate effects are fixed (Martinussen and Scheike, 2009). However in our setting covariate dependence may vary arbitrarily with time, so any selection criteria must be defined and applied locally. This paper adapts two existing selection methods from classical linear regression: generalized cross-validation and the L-curve.

2. Singular value decomposition

Whatever pseudo-inverse we choose, it is instructive to consider the singular value decomposition (SVD) of $\mathbf{Y}(t)$:

$$\mathbf{Y}(t) = \mathbf{U}(t)\mathbf{D}(t)\mathbf{V}(t)^\top, \tag{2.1}$$

where $\mathbf{U}(t)$ is an $n\times n$ matrix such that $\mathbf{U}(t)^\top\mathbf{U}(t) = \mathbf{I}$, $\mathbf{V}(t)$ is a $(p+1)\times(p+1)$ matrix satisfying $\mathbf{V}(t)^\top\mathbf{V}(t) = \mathbf{V}(t)\mathbf{V}(t)^\top = \mathbf{I}$, and $\mathbf{D}(t)$ is an $n\times(p+1)$ diagonal matrix of singular values $d_1, \ldots, d_{p+1}$. The generalized inverse for ridge regression (1.2) is thus equivalent to

$$\mathbf{Y}(t)^-_{\text{RR}} = \left(\mathbf{V}(t)\mathbf{D}(t)^2\mathbf{V}(t)^\top + \eta(t)\mathbf{I}\right)^{-1}\mathbf{V}(t)\mathbf{D}(t)\mathbf{U}(t)^\top = \mathbf{V}(t)\mathbf{F}(t)_{\text{RR}}\mathbf{U}(t)^\top,$$

where $\mathbf{F}(t)_{\text{RR}} = (\mathbf{D}(t)^2 + \eta(t)\mathbf{I})^{-1}\mathbf{D}(t)$.

When $\text{rank}(\mathbf{Y}(t)) = q < p+1$, $d_{q+1}, \ldots, d_{p+1}$ are zero and (2.1) may then be rewritten as

$$\mathbf{Y}(t) = \mathbf{U}(t)_q \text{diag}(d_1, \ldots, d_q)\mathbf{V}(t)_q^\top,$$

where the subscript q denotes the submatrix given by the first q columns of the original matrix. Brown (1978) used this form of truncated SVD to obtain an OLS solution under rank deficiency for classical linear regression. So, although our focus in this paper is on ridge regression, we should note that singularities may also be addressed with use of the pseudo-inverse

$$\mathbf{Y}(t)^-_{\text{TSVD}} = \mathbf{V}(t)_q \text{diag}(1/d_1, \ldots, 1/d_q)\mathbf{U}(t)_q^\top. \tag{2.2}$$

3. Ridge parameter selection

From the outset, strategies for ridge parameter selection have aimed to minimize the mean square error (MSE; Hoerl and Kennard, 1970). Suppressing the dependence on time, the increment $\Delta\hat{\mathbf{A}}_{\text{RR}}(t)$ has MSE

$$\begin{aligned}
\text{MSE}\left(\Delta\hat{\mathbf{A}}_{\text{RR}}\right) &= \mathbf{V}\text{MSE}\left(\mathbf{F}_{\text{RR}}\mathbf{U}^\top\Delta\mathbf{N}\right)\mathbf{V}^\top \\
&= \mathbf{V}\left[\text{tr}\left\{\text{Var}\left(\mathbf{F}_{\text{RR}}\mathbf{U}^\top\Delta\mathbf{N}\right)\right\} + \left\{(\mathbf{I}-\mathbf{F}_{\text{RR}})\mathbf{U}^\top\Delta\mathbf{N}\right\}^{\otimes 2}\right]\mathbf{V}^\top \\
&= \mathbf{V}\left\{\sum_{j=1}^{p+1}\frac{d_j^2\mathbf{u}_j^\top\text{Var}(\Delta\mathbf{N})\mathbf{u}_j}{\left(d_j^2+\eta\right)^2} + \sum_{j=1}^{p+1}\frac{\eta^2\left(\mathbf{u}_j^\top\Delta\mathbf{N}\right)^2}{\left(d_j^2+\eta\right)^2}\right\}\mathbf{V}^\top,
\end{aligned}$$

where $\mathbf{u}_j$ is the jth column of $\mathbf{U}$ ($j = 0, \ldots, p$). Assuming that $\mathbf{N} - \int \text{d}\,\mathbf{A}$ is a martingale and from Andersen et al. (1993, Proposition II.4.1), $\text{Var}(\Delta\mathbf{N})$ is consistently estimated by $\text{diag}(\Delta\mathbf{N})$. The $\eta(t)$ that minimizes the MSE is then approximately equal to one—a value that would have undue influence on the model fit. So we must consider alternatives for the minimum MSE criterion in order to achieve a reasonable bias–variance trade-off.

3.1. The L-curve

The L-curve is a plot of the regularized solution norm versus the corresponding residual norm, both in log scale (Hansen, 1992). It displays the trade-off between the size of the estimate and the fit to the data. The L-curve criterion for selecting the ridge parameter is to balance these quantities by selecting the parameter value at which the L-curve reaches maximum curvature.

In our setting, the solution norm for the increment in the cumulative regression functions is

$$s(\eta) = \|\Delta\hat{\mathbf{A}}_{\text{RR}}\|^2 = \|\mathbf{V}\mathbf{F}_{\text{RR}}\mathbf{U}^\top\Delta\mathbf{N}\|^2 = \sum_{j=1}^{p+1}\frac{d_j^2(\mathbf{u}_j^\top\Delta\mathbf{N})^2}{(d_j^2+\eta)^2},$$

where we suppress the dependence on time for brevity. The corresponding residual norm is the residual sum of squares

$$r(\eta) = \|\Delta\mathbf{N} - \mathbf{Y}\Delta\hat{\mathbf{A}}_{\text{RR}}\|^2 = 1 - 2\sum_{j=1}^{p+1}\frac{d_j^2(\mathbf{u}_j^\top\Delta\mathbf{N})^2}{d_j^2+\eta} + \sum_{j=1}^{p+1}\frac{d_j^4(\mathbf{u}_j^\top\Delta\mathbf{N})^2}{(d_j^2+\eta)^2}.$$