dgLARS method for relative risk regression models

Il metodo dgLARS per i modelli di regressione del rischio relativo

Luigi Augugliaro and Angelo M. Mineo

Abstract With the introduction of high-throughput technologies in clinical and epidemiological studies, the need for inferential tools that are able to deal with fat data-structures, i.e., relatively small number of observations compared to the number of features, is becoming more prominent. To solve this problem, in this paper we propose an extension of the dgLARS method to relative risk regression model. The main idea of proposed method is to use the differential geometric structure of the partial likelihood function in order to select the optimal subset of covariates.

Abstract *L'introduzione di tecnologie di screening ad elevata capacità negli studi clinici ed epidemiologici ha reso preminente il problema dello sviluppo di metodologie inferenziali applicabili ai casi in cui la numerosità campionaria è inferiore al numero di parametri. In questo lavoro proponiamo un'estensione del metodo dgLARS ai modello di regressione del rischio relativo. L'idea di fondo del metodo proposto è quella di utilizzare la struttura geometrica della partial likelihood al fine di selezionare il sottoinsieme ottimo di variabili esplicative.*

Key words: dgLARS, relative risk regression models, sparsity, survival analysis.

1 Introduction

In the study of the dependence of survival time on covariates, the Cox proportional hazards model [3] has proved to be a major tool in many clinical and epidemiological applications. However, when the number of features is large, the simple Cox

Luigi Augugliaro
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: luigi.augugliaro@unipa.it

Angelo M. Mineo
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: angelo.mineo@unipa.it

proportional breaks down. Many variable selection techniques for linear regression models have been extended to the context of survival models. They include best-subset selection, stepwise selection, asymptotic procedures based on score tests, Wald tests and other approximate chi-squared testing procedures, bootstrap procedures and Bayesian variable selection. However, the theoretical properties of these methods are generally unknown. Recently a family of penalized partial likelihood methods, such as the Lasso [11] and the smoothly clipped absolute deviation method [5] were proposed for the Cox proportional hazards model. By shrinking some regression coefficients to zero, these methods select important variables and estimate the regression model simultaneously. Whereas the Lasso estimator does not possess oracle properties, the smoothly clipped absolute deviation estimator for linear models, has better theoretical properties. However, the non-convex form of the penalty term of the latter makes its optimization challenging in practice, and the solutions may suffer from numerical instability. In this paper we propose an alternative to the penalized inference methods. We extend the differential-geometric least angle regression method (dgLARS) [1] to the case of the Cox proportional hazards model.

2 The differential geometrical structure of a relative risk regression model

In analyzing survival data, one of the most important tools is the hazard function. Formally, let T be the absolutely continuous random variable associated with the survival time and let $f(t)$ be the corresponding probability density function. The hazard function is defined as $\lambda(t) = f(t)/\{1-\int_0^t f(s)ds\}$ and specifies the instantaneous rate at which failures occur for subjects that are surviving at time t. Suppose that the hazard function $\lambda(t)$ can depend on a p-dimensional vector of covariates which can depend on time and denoted by $\boldsymbol{x}(t) = (x_1(t),\ldots,x_p(t))^\top$. The relative risk regression models [10] are based on the assumption that the vector $\boldsymbol{x}(t)$ influence the hazard function $\lambda(t)$ by the following relation $\lambda(t;\boldsymbol{x}) = \lambda_0(t)\psi(\boldsymbol{x}(t);\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a p-dimensional vector of unknown fixed parameters, $\lambda_0(t)$ is the base hazard function at time t which is left unspecified and, finally, $\psi : \mathbb{R} \to \mathbb{R}$ is a fixed twice continuously differentiable function, called the relative risk function. The parameter space is such that $\psi(\boldsymbol{x}(t);\boldsymbol{\beta}) > 0$ for each $\boldsymbol{\beta}$; we also assume that the relative risk function is normalized, i.e., $\psi(\boldsymbol{0};\boldsymbol{\beta}) = 0$.

Suppose that n observations are available and let with t_i the ith observed failure time. Assume that we have k uncensored failure times and let us denoted by D the set of indices for which the corresponding failure time is observed. The remaining failure times are right censored. Under the assumption of independent censoring, the inference about $\boldsymbol{\beta}$ can be carried out by the following partial likelihood function

$$\mathscr{L}_p(\boldsymbol{\beta}) = \prod_{i\in D} \frac{\psi(\boldsymbol{x}_i(t_i);\boldsymbol{\beta})}{\sum_{j\in R(t_i)} \psi(\boldsymbol{x}_j(t_i);\boldsymbol{\beta})}, \tag{1}$$

where $R(t)$ denotes the risk set, i.e. the set of indices corresponding to the subjects how have not failed and are still under observation just prior to time t. In order to extend the dgLARS method to the relative risk regression model, it is useful to see the partial likelihood (1) as arising from a multinomial sampling scheme. Consider an index $i \in D$ and let $\boldsymbol{Y}_i = (Y_{ih})_{h \in R(t_i)}$ be a multinomial random variable with sample size equal to 1 and cell probabilities $\boldsymbol{\pi}_i = (\pi_{ih})_{h \in R(t_i)} \in \Pi_i$. Assuming that the random vectors $\boldsymbol{Y}_i$ are independent, the joint probability density function is an element of the set $S = \left\{ \prod_{i \in D} \prod_{h \in R(t_i)} \pi_{ih}^{y_{ih}} : (\boldsymbol{\pi}_i)_{i \in D} \in \bigotimes_{i \in D} \Pi_i \right\}$. In the following of our differential geometric constructions, the set S will play the role of ambient space. Consider the following model for the conditional expected value of the random variable Y_{ih}: $E_{\boldsymbol{\beta}}(Y_{ih}) = \pi_{ih}(\boldsymbol{\beta}) = \psi(\boldsymbol{x}_h(t_i); \boldsymbol{\beta}) / \sum_{j \in R(t_i)} \psi(\boldsymbol{x}_j(t_i); \boldsymbol{\beta})$, then our model space is the set $M = \left\{ \prod_{i \in D} \prod_{h \in R(t_i)} \pi_{ih}(\boldsymbol{\beta})^{y_{ih}} : (\boldsymbol{\pi}_i)_{i \in D} \in \bigotimes_{i \in D} \Pi_i \right\}$. The partial likelihood (1) is formally equivalent to the likelihood function associated with the model space M if we assume that for each $i \in D$, the observed y_{ih} is equal to one if h is equal to i and zero otherwise. Let $\ell(\boldsymbol{\beta})$ be the log-likelihood function associated to the model space M and let $\partial_m \ell(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\beta}) / \partial \beta_m$. The tangent space $T_{\boldsymbol{\beta}} M$ of M at the model point $\prod_{i \in D} \prod_{h \in R(t_i)} \pi_{ih}(\boldsymbol{\beta})^{y_{ih}}$ is defined as that linear vector space spanned by the p elements of the score vector, formally $T_{\boldsymbol{\beta}} M = span\{\partial_1 \ell(\boldsymbol{\beta}), \ldots, \partial_p \ell(\boldsymbol{\beta})\}$. Under the standard regularity conditions, it is easy to see that $T_{\boldsymbol{\beta}} M$ is the linear vector space of the random variables $v(\boldsymbol{\beta}) = \sum_{m=1}^{p} v_m \partial_m \ell(\boldsymbol{\beta})$ with zero expected value and finite variance. As a simple consequence of the chain rule we have the following identity for any tangent vector belonging to the tangent space $T_{\boldsymbol{\beta}} M$, i.e.

$$v(\boldsymbol{\beta}) = \sum_{m=1}^{p} v_m \partial_m \ell(\boldsymbol{\beta}) = \sum_{i \in D} \sum_{h \in R(t_i)} \left(\sum_{m=1}^{p} v_m \frac{\partial \pi_{ih}(\boldsymbol{\beta})}{\partial \beta_m} \right) \frac{\partial \ell(\boldsymbol{\beta})}{\partial \pi_{ih}} = \sum_{i \in D} \sum_{h \in R(t_i)} w_{ih} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \pi_{ih}},$$

which shows that $T_{\boldsymbol{\beta}} M$ is a linear vector subspace of the tangent space $T_{\boldsymbol{\beta}} S$ spanned by the random variables $\partial_{ih} \ell(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\beta}) / \partial \pi_{ih}$. To define the notion of angle between two given tangent vectors belonging to $T_{\boldsymbol{\beta}} M$, say $v(\boldsymbol{\beta})$ and $w(\boldsymbol{\beta})$, we shall use the information metric [9], i.e.

$$\langle v(\boldsymbol{\beta}); w(\boldsymbol{\beta}) \rangle_{\boldsymbol{\beta}} = E_{\boldsymbol{\beta}}(v(\boldsymbol{\beta}) w(\boldsymbol{\beta})) = \boldsymbol{v}^{\top} I(\boldsymbol{\beta}) \boldsymbol{w}, \tag{2}$$

where $\boldsymbol{v} = (v_1 \ldots, v_p)^{\top}$, $\boldsymbol{w} = (w_1 \ldots, w_p)^{\top}$ and $I(\boldsymbol{\beta})$ is the Fisher information matrix evaluated at $\boldsymbol{\beta}$. As observed in [6], the matrix $I(\boldsymbol{\beta})$ used in (2) is not exactly equal to the Fisher information matrix of the relative risk regression model, however it has the appropriate asymptotic properties for the inference [8].

3 dgLARS method for relative risk regression model

dgLARS method is a sequential method developed to estimate a sparse solution curve embedded in the parameter space based on a differential geometric characterization of the Rao score test statistic obtained considering the inner prod-

uct between the bases of the tangent space $T_{\boldsymbol{\beta}}M$ and the tangent residual vector $r(\boldsymbol{\beta}) = \sum_{i \in D} \sum_{h \in R(t_i)} r_{ih}(\boldsymbol{\beta}) \partial_{ih} \ell(\boldsymbol{\beta}) \in T_{\boldsymbol{\beta}} S$, where $r_{ih}(\boldsymbol{\beta}) = y_{ih} - \pi_{ih}(\boldsymbol{\beta})$. As observed in [1], the mth signed Rao score test statistic satisfies the following differential geometric characterization, i.e.

$$r_m^u(\boldsymbol{\beta}) = I_{mm}^{-1/2}(\boldsymbol{\beta}) \partial_m \ell(\boldsymbol{\beta}) = \cos(\rho_m(\boldsymbol{\beta})) \| r(\boldsymbol{\beta}) \|_{\boldsymbol{\beta}}, \tag{3}$$

where $I_{mm}(\boldsymbol{\beta})$ is the Fisher information for β_m, $\| r(\boldsymbol{\beta}) \|_{\boldsymbol{\beta}}^2 = E_{\boldsymbol{\beta}}(r(\boldsymbol{\beta})^2)$ and $\cos(\rho_m(\boldsymbol{\beta}))$ is a generalization of the Euclidean notion of angle between the mth column of the design matrix and the residual vector. Characterization (3) gives us a natural way to generalize the equiangularity condition [4]: two given predictors, say the mth and nth, satisfy the generalizes equiangularity condition at the point $\boldsymbol{\beta}$ when $|r_m^u(\boldsymbol{\beta})| = |r_n^u(\boldsymbol{\beta})|$. Inside the dgLARS theory, the generalized equiangularity condition is used to identify the predictors that are included in the model.

The nonzero estimates are formally defined as follows. For any data set there is a finite sequence of transition points, say $\gamma^{(1)} \geq \ldots \geq \gamma^{(K)} \geq 0$, such that for any fixed γ between $\gamma^{(k+1)}$ and $\gamma^{(k)}$ the sub vector of the non nonzero dgLARS estimates, denoted as $\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma) = (\hat{\beta}_m(\gamma))_{m \in \hat{\mathscr{A}}}$, satisfies the following conditions:

$$\begin{aligned} r_m^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma)\} &= s_m \gamma, \quad m \in \hat{\mathscr{A}} \\ |r_n^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma)\}| &< \gamma, \quad n \notin \hat{\mathscr{A}} \end{aligned}$$

where $s_m = \operatorname{sign}\{\hat{\beta}_m(\gamma)\}$ and $\hat{\mathscr{A}} = \{m : \hat{\beta}_m(\gamma) \neq 0\}$, called active set, is the set of the indices of the predictors that are included in the current model, called active predictors. In any transition point, say for example $\gamma^{(k)}$, one of the following two conditions occurs:

1. there is a non active predictor, say the nth, satisfying the generalized equiangularity condition with any active predictor, i.e.,

$$|r_n^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma^{(k)})\}| = |r_m^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma^{(k)})\}| = \gamma^{(k)}, \tag{4}$$

   for any m in $\hat{\mathscr{A}}$, then it is included in the active set;
2. there is an active predictor, say the mth, such that

$$\operatorname{sign}[r_m^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma^{(k)})\}] \neq \operatorname{sign}\{\hat{\beta}_m(\gamma^{(k)})\}, \tag{5}$$

   then it is removed from the active set.

Given the previous definition, the path of solutions can be constructed in the following way. Since we are working with a class of regression models without intercept term, the starting point of the dgLARS curve is the zero vector this means that, at the starting point, the p predictors are ranked using $|r_m^u(\mathbf{0})|$. Suppose that $a_1 = \arg\max_m |r_m^u(\mathbf{0})|$, then $\hat{\mathscr{A}} = \{a_1\}$, $\gamma^{(1)}$ is set equal to $|r_{a_1}^u(\mathbf{0})|$ and the first segment of the dgLARS curve is implicitly defined by the nonlinear equation $r_{a_1}^u\{\hat{\beta}_{a_1}(\gamma)\} - s_{a_1}\gamma = 0$. The proposed method traces the first segment of the

dgLARS curve reducing γ until we find the transition point $\gamma^{(2)}$ corresponding to the inclusion of a new index in the active set, in other words, there exists a predictor, say the a_2th, satisfying condition (4), then a_2 is included in $\hat{\mathscr{A}}$ and the new segment of the dgLARS curve is implicitly defined by the system with nonlinear equations:

$$r^{u}_{a_i}\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma)\} - s_{a_i}\gamma = 0, \quad a_i \in \hat{\mathscr{A}},$$

where $\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma) = (\hat{\beta}_{a_1}(\gamma), \hat{\beta}_{a_2}(\gamma))^{\top}$. The second segment is computed reducing γ and solving the previous system until we find the transition point $\gamma^{(3)}$. At this point, if condition (4) occurs a new index is included in $\hat{\mathscr{A}}$ otherwise condition (5) occurs and an index is removed from $\hat{\mathscr{A}}$. In the first case the previous system is updated adding a new nonlinear equation while, in the second case, a nonlinear equation is removed. The curve is traced as previously described until parameter γ is equal to a fixed value that can be zero, if the the sample size is large enough, or a positive value if we are working in a high-dimensional setting, i.e., the number of predictors is larger than the sample size. In this way we can avoid the problems coming from the overfitting of the model. From a computational point of view, the entire dgLARS curve can be computed using the algorithms proposed in [2, 7].

4 Simulation study

In this section we compare the method introduced in Section 3 with three popular algorithms named CoxNet, CoxPath, and CoxPen. Given the fact that these methods have only been implemented only for Cox regression model, our comparison will focus on this kind of relative risk regression model. In the following of this section, dgLARS method applied to the Cox regression model is named dgCox model.

We simulated one hundred datasets from a Cox regression model where the survival times t_i ($i = 1, \ldots, n$) follow an exponential distributions with parameter $\lambda_i = \exp(\boldsymbol{\beta}^{\top}\boldsymbol{x}_i)$, and $\boldsymbol{x}_i$ is sampled from a p-variate normal distribution $\mathrm{N}(\mathbf{0}, \Sigma)$; the entries of Σ are fixed to $\mathrm{corr}(X_m, X_n) = \rho^{|m-n|}$ with $\rho = 0.9$. The censorship is randomly assigned to the survival times with probability $\pi \in \{0.2, 0.4\}$. To emulate an high-dimensional setting, we fixed the sample size to 50, the number of predictors to 100 and $\beta_m = 0.5$ ($m = 1, \ldots, 30$); the remaining regression coefficients are zero in order to have a sparse vector. To remove the effects coming from the information measure used to select the optimal point of each paths of solutions, we evaluated the global behaviour of the paths by using the ROC curve and the corresponding Area Under the Curve (AUC). Figure 1 shows that dgCox model is clearly the superior approach for both levels of censorship. For the same false positive rate, the true positive rate of the dgCox method is around 10% higher than the rate obtained by CoxNet, CoxPath and CoxPen.

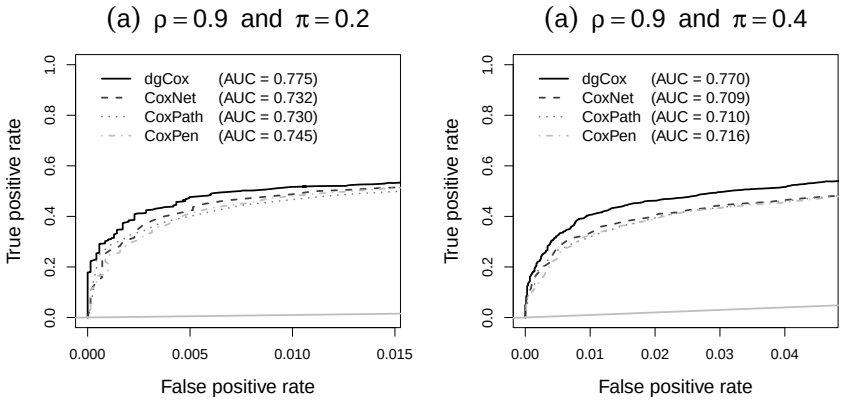

Fig. 1 Results from the simulation study; for each scenario we show the averaged ROC curve for dgCox, CoxNet, CoxPath and CoxPen algorithm. The average Area Under the Curve (AUC) is also reported. The 45-degree diagonal is also included in the plots.

References

1. Augugliaro L., Mineo, A. M., Wit, E.: Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models. J. Roy. Statist. Soc. Ser. B. **75**(3), 471–498 (2013)
2. Augugliaro L., Mineo, A. M., Wit, E.: dglars: An R package to estimate sparse generalized linear models. J. Stat. Soft. **59**(8), 1–40 (2014)
3. Cox, D.: Regression models and life-tables. J. Roy. Statist. Soc. Ser. B. **34**(2), 187–220 (1972)
4. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Statist. **32**(2), 407–499 (2004)
5. Fan, J., Li, R.: Variable selection for Cox?s proportional hazards model and frailty model. Ann. Statist. **30**(1), 74–99 (2002)
6. Moolgavkar, S., Venzon, D. J.: Confidence regions in curved exponential families: application to matched case-control and survival studies with general relative risk function. Ann. Statist. **15**(1), 346–359 (1987)
7. Pazira, H., Augugliaro, L., Wit, E. C.: Extended differential geometric lars for high-dimensional glms with general dispersion parameter. Stat. Comput. **28**(4), 753–774 (2018)
8. Prentice, R., Self, S.: Asymptotic distribution theory for Cox-type regression models with general relative risk form. Ann. Statist. **11**(3), 804–813 (1983)
9. Rao, C. R.: On the distance between two populations. Sankhyā. **9**, 246–248 (1949)
10. Thomas, D.: General relative-risk models for survival time and matched case-control analysis. Biometrika. **37**(4), 673–686 (1981)
11. Tibshirani, R.: The lasso method for variable selection in the Cox model. Stat. Med. **16**, 385–395 (1997)