Biostatistics (2020) **21**, 2, *pp.* e131–e147
doi:10.1093/biostatistics/kxy060
Advance Access publication on October 30, 2018

Sparse relative risk regression models

ERNST C. WIT*

Institute of Computational Science, USI, Via Buffi 13, 6900 Lugano, Switzerland
wite@usi.ch

LUIGI AUGUGLIARO
Department of Economics, Business and Statistics, University of Palermo, Building 13, Viale delle Scienze, 90128 Palermo, Italy

HASSAN PAZIRA
Bernoulli Institute, University of Groningen, Nijenborg 9, 9747 AG Groningen, The Netherlands

JAVIER GONZÁLEZ†
Amazon Research Cambridge, Poseidon House, Castle Park, Cambridge, UK

FENTAW ABEGAZ
Bernoulli Institute, University of Groningen, Nijenborg 9, 9747 AG Groningen, The Netherlands and Department of Pediatrics and Systems Biology Centre for Energy Metabolism and Ageing, University of Groningen, University Medical Center Groningen, 9700 AD Groningen, The Netherlands

Summary

Clinical studies where patients are routinely screened for many genomic features are becoming more routine. In principle, this holds the promise of being able to find genomic signatures for a particular disease. In particular, cancer survival is thought to be closely linked to the genomic constitution of the tumor. Discovering such signatures will be useful in the diagnosis of the patient, may be used for treatment decisions and, perhaps, even the development of new treatments. However, genomic data are typically noisy and high-dimensional, not rarely outstripping the number of patients included in the study. Regularized survival models have been proposed to deal with such scenarios. These methods typically induce sparsity by means of a coincidental match of the geometry of the convex likelihood and a (near) non-convex regularizer. The disadvantages of such methods are that they are typically non-invariant to scale changes of the covariates, they struggle with highly correlated covariates, and they have a practical problem of determining the amount of regularization. In this article, we propose an extension of the differential geometric least angle regression method for sparse inference in relative risk regression models. A software implementation of our method is available on github (https://github.com/LuigiAugugliaro/dgcox).

Keywords: dgLARS; Gene expression data; High-dimensional data; Relative risk regression models; Sparsity; Survival analysis.

*To whom correspondence should be addressed.
†The work in this paper was done before joining Amazon.

1. INTRODUCTION

Advances in genomic technologies have meant that many new clinical studies in cancer survival include a variety of genomic measurements, ranging from gene expression to SNP data. Studying the relationship between survival and genomic markers can be useful for a variety of reasons. If a genomic signature can be found, then patients can be given more accurate survival information. Furthermore, treatment and care may be adjusted to the prospects of an individual patient. Eventually, the genomic signature combined with information from other studies may be used to identify drug targets. We will focus on four recent studies of cancer survival for four different tumors. Our aim is to find a reproducible sparse predictor for cancer survival.

Sparse inference in the past two decades has been dominated by methods that penalize typically convex likelihoods by functions of the parameters that happen to induce solutions with many zeros. The Lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), l_0 (Rippe *and others*, 2012), and the SCAD (Fan and Li, 2001) penalties are examples of such penalties that, depending on some tuning parameter, conveniently shrink estimates to exact zeros. Also in survival analysis, these methods have been introduced. Tibshirani (1997) applied the Lasso penalty to the Cox proportional hazards model. Gui and Li (2005), Sohn *and others* (2009), and Goeman (2010) suggested important computational improvements to make the calculation of the Lasso estimator in the Cox proportional hazards model more efficient. Although the Lasso penalty induces sparsity, it is well known to suffer from possible inconsistent selection of variables.

Whereas penalized inference is convenient, justification of the penalty is somewhat problematic. Interpreting the solution as a Bayesian MAP estimator with a particular prior on the parameters seems to merely reformulate the problem, rather than solving it. Furthermore, the methods suffer from being not invariant under scale transformations of the explanatory variables. This means that measuring, e.g., height in centimeters or inches can and probably will result in dramatically different answers. Therefore, most penalized regression methods start their exposition by assuming that the variables are appropriately renormalized. This is clearly a merely algorithmic device and simply begs the question of invariance. Clearly the strongest argument in favor of some of these methods are their asymptotic properties. Nevertheless, what this means in the small sample settings encountered in practice is also problematic.

In this article, we will approach sparsity directly from a likelihood point of view. The angle between the covariates and the tangent residual vector within the likelihood manifold provides a direct and scale-invariant way to assess the importance of the individual covariates. The idea is similar to the least angle regression approach proposed by Efron *and others* (2004). However, rather than using it as a computational device for obtaining Lasso solutions, we view the method in its own right as in Augugliaro *and others* (2013). Moreover, the method extends directly beyond the Cox proportional hazard model. In fact, we will focus on general relative risk survival models.

The remaining part of this article is structured as follows. In Section 2, we introduce the relative risk regression model and in Section 3, first we derive the differential geometric structure of a relative regression model and then we use it to extend the differential geometric least angle regression (dgLARS) method (Augugliaro *and others*, 2013). In this section, by appealing to the theory of Z-estimation, we derive a robust way of selecting a unique point of the path of solutions defined by dgLARS. In Section 4, by simulation studies, we compare the performance of the proposed method to other sparse survival regression approaches, especially in the presence of correlated predictors. In Section 5, we return to the motivating cancer survival studies and employ differential geometric Cox proportional hazards modelling to find a genetic signature for cancer survival in skin, colon, prostate, and ovarian cancer. Finally, in Section 6 we draw some conclusions.

2. Relative risk regression models

In analyzing survival data, one of the most important tools is the hazard function, which is used to express the risk or hazard of failure at some time t. 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 $\lambda(t) = f(t)/\{1 - \int_0^t f(s)\mathrm{d}s\}$ and specifies the instantaneous rate at which failures occur for subjects that are surviving at time t. Suppose that the hazard function can depend on a p-dimensional, possibly time-dependent, vector of covariates, denoted by $\mathbf{x}(t) = (x_1(t), \ldots, x_p(t))^\top$. Relative risk regression models are based on the assumption that $\mathbf{x}(t)$ influences the hazard function through the following relation

$$\lambda(t; \mathbf{x}) = \lambda_0(t)\psi\{\mathbf{x}(t); \boldsymbol{\beta}\}, \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathcal{B} \subseteq \mathbb{R}^p$ is a p-dimensional vector of unknown fixed parameters and $\lambda_0(t)$ is the base line hazard function at time t, which is left unspecified. Finally, $\psi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a differentiable function, called the *relative risk function*, and the parameter space $\mathcal{B}$ is such that $\psi\{\mathbf{x}(t); \boldsymbol{\beta}\} > 0$ for each $\boldsymbol{\beta} \in \mathcal{B}$. We also assume that the relative risk function is normalized, i.e., $\psi(\mathbf{0}; \boldsymbol{\beta}) = 1$. Model (2.1), originally proposed in Thomas (1981), clearly extends the usual Cox regression model (Cox, 1972) which is obtained when $\psi\{\mathbf{x}(t); \boldsymbol{\beta}\} = \exp\{\boldsymbol{\beta}^\top \mathbf{x}(t)\}$, and allows us to work with applications in which the exponential form of the relative risk function is not the best choice. This issue was observed in Oakes (1981) and further underlined in Cox (1981). As a motivating example for the generalization (2.1), several authors have noted that the linear relative risk function $\psi\{\mathbf{x}(t); \boldsymbol{\beta}\} = 1 + \boldsymbol{\beta}^\top \mathbf{x}(t)$ provides a natural framework within which to assess departures from an additive relative risk model when two or more risk factors are studied in relation to the incidence of a disease (see e.g., Thomas 1981; Prentice *and others* 1983; Prentice and Mason 1996, among the other). Other possible choices for the relative risk functions are the logit relative risk function $\psi\{\mathbf{x}(t); \boldsymbol{\beta}\} = \log[1 + \exp\{\boldsymbol{\beta}^\top \mathbf{x}(t)\}]$, proposed by Efron (1977), or the the excess relative risk model $\psi\{\mathbf{x}(t); \boldsymbol{\beta}\} = \prod_{m=1}^p \{1 + x_m(t)\beta_m\}$.

Suppose that n observations are available and let t_i be the ith observed failure time. Assume that we have k uncensored and untied failure times and let $\mathcal{D}$ be the set of indices for which the corresponding failure time is observed; the remaining failure times are right censored. As explained in Cox and Oakes (1984), if we denote by $\mathcal{R}(t)$ the risk set, i.e., the set of indices corresponding to the subjects who have not failed and are still under observation just prior to time t, under the assumption of independent censoring, inference about the $\boldsymbol{\beta}$ can be carried out by the partial likelihood function

$$\mathcal{L}_p(\boldsymbol{\beta}) = \prod_{i \in \mathcal{D}} \frac{\psi\{\mathbf{x}_i(t_i); \boldsymbol{\beta}\}}{\sum_{j \in \mathcal{R}(t_i)} \psi\{\mathbf{x}_j(t_i); \boldsymbol{\beta}\}}. \tag{2.2}$$

When the exponential relative risk function is used in model (2.1) and we work with fixed covariates, (2.2) is clearly equal to the original partial likelihood introduced in Cox (1972) and discussed in great detail in Cox (1975).

3. Sparse relative risk regression

In this section, we extend the dgLARS method (Augugliaro *and others*, 2013) to the relative risk regression models. The basic idea underlying the dgLARS method is to use the differential geometrical structure of a Generalized Linear Model (GLM) (McCullagh and Nelder, 1989) to generalize the LARS method (Efron *and others*, 2004). This means that, our first step is relate the partial likelihood (2.2) with the likelihood function of a specific GLM. As originally observed in Thomas (1977), and studied in greater detail in

Prentice and Breslow (1978), to solve this problem we shall use the identity existing between the partial likelihood and the likelihood function of a logistic regression model for matched case–control studies. The idea to use this identity to study the differential geometrical structure of a relative risk regression model is not new and was originally used in Moolgavkar and Venzon (1987) to construct approximated confidence regions for the proportional hazards model.

3.1. *Differential geometrical structure of the relative risk regression model*

The partial likelihood (2.2) can be seen as arising from a multinomial sample scheme. Consider an index $i \in \mathcal{D}$ and let $\mathbf{Y}_i = (Y_{ih})_{h \in \mathcal{R}(t_i)}$ be a multinomial random variable with sample size equal to 1 and cell probabilities $\boldsymbol{\pi}_i = (\pi_{ih})_{h \in \mathcal{R}(t_i)} \in \Pi_i$, i.e., $p(\mathbf{y}; \boldsymbol{\pi}_i) = \prod_{h \in \mathcal{R}(t_i)} \pi_{ih}^{y_{ih}}$. Assuming that the random vectors $\mathbf{Y}_i$ are independent, the joint probability density function is an element of the set $\mathcal{S} = \left\{ \prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \pi_{ih}^{y_{ih}} : (\boldsymbol{\pi}_i)_{i \in \mathcal{D}} \in \bigotimes_{i \in \mathcal{D}} \Pi_i \right\}$, called the ambient space. We would like to underline that our differential geometric constructions are invariant to the chosen parameterization, which means that $\mathcal{S}$ can be equivalently defined by the canonical parameter vector and this will not change the results. In this article, we prefer to use the mean value parameter vector to specify our differential geometrical description because this will make the relationship with the partial likelihood (2.2) clearer. If we model the expected value of Y_{ih} as follows:

$$E_{\boldsymbol{\beta}}(Y_{ih}) = \pi_{ih}(\boldsymbol{\beta}) = \frac{\psi\{\mathbf{x}_h(t_i); \boldsymbol{\beta}\}}{\sum_{j \in \mathcal{R}(t_i)} \psi\{\mathbf{x}_j(t_i); \boldsymbol{\beta}\}}, \tag{3.1}$$

it is easy to see that the partial likelihood (2.2) is formally equivalent to the likelihood function associated with the model space $\mathcal{M} = \left\{ \prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \{\pi_{ih}(\boldsymbol{\beta})\}^{y_{ih}} : \boldsymbol{\beta} \in \mathcal{B} \right\}$ assuming that for each $i \in \mathcal{D}$, the observed y_{ih} is equal to one if h is equal to i and zero otherwise. Let $\ell(\boldsymbol{\beta}) = \sum_{i \in \mathcal{D}} \sum_{h \in \mathcal{R}(t_i)} Y_{ih} \log \pi_{ih}(\boldsymbol{\beta})$ be the log-likelihood function associated to the model space $\mathcal{M}$ and let $\partial_m \ell(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\beta}) / \partial \beta_m$. The tangent space $T_{\boldsymbol{\beta}}\mathcal{M}$ of $\mathcal{M}$ at the model point $\prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \{\pi_{ih}(\boldsymbol{\beta})\}^{y_{ih}}$ is defined as the linear vector space spanned by the p elements of the score vector, formally, $T_{\boldsymbol{\beta}}\mathcal{M} = \text{span}\{\partial_1 \ell(\boldsymbol{\beta}), \ldots, \partial_p \ell(\boldsymbol{\beta})\}$. Under standard regularity conditions, it is easy to see that $T_{\boldsymbol{\beta}}\mathcal{M}$ 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 expectation and finite variance. Applying the chain rule, for any tangent vector belonging to $T_{\boldsymbol{\beta}}\mathcal{M}$ we have that

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

where $\partial_{ih} \ell(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\beta}) / \partial \pi_{ih}$; the previous expression shows that $T_{\boldsymbol{\beta}}\mathcal{M}$ is a linear sub vector space of the tangent space $T_{\boldsymbol{\beta}}\mathcal{S}$ spanned by the random variables $\partial_{ih} \ell(\boldsymbol{\beta})$. To define the notion of angle between two given tangent vectors belonging to $T_{\boldsymbol{\beta}}\mathcal{M}$, say $v_{\boldsymbol{\beta}} = \sum_{m=1}^{p} v_m \partial_m \ell(\boldsymbol{\beta})$ and $w_{\boldsymbol{\beta}} = \sum_{n=1}^{p} w_n \partial_n \ell(\boldsymbol{\beta})$, we shall use the information metric (Rao, 1949), i.e.,

$$\langle v_{\boldsymbol{\beta}}; w_{\boldsymbol{\beta}} \rangle_{\boldsymbol{\beta}} = E_{\boldsymbol{\beta}}(v_{\boldsymbol{\beta}} w_{\boldsymbol{\beta}}) = \sum_{m,n=1}^{p} E_{\boldsymbol{\beta}}\{\partial_m \ell(\boldsymbol{\beta}) \partial_n \ell(\boldsymbol{\beta})\} v_m w_n = \mathbf{v}^{\top} I(\boldsymbol{\beta}) \mathbf{w}, \tag{3.2}$$

where $\mathbf{v} = (v_1, \ldots, v_p)^{\top}$, $\mathbf{w} = (w_1, \ldots, w_p)^{\top}$ and $I(\boldsymbol{\beta})$ is the Fisher information matrix evaluated at $\boldsymbol{\beta}$. As observed in Moolgavkar and Venzon (1987), the matrix $I(\boldsymbol{\beta})$ used in (3.2) is not exactly equal to the Fisher

information matrix of the relative risk regression model, however it has the same asymptotic properties for inference. Finally, to complete our differential geometric framework we need to introduce the tangent residual vector $r_{\boldsymbol{\beta}} = \sum_{i\in\mathcal{D}}\sum_{h\in\mathcal{R}(t_i)} r_{ih}(\boldsymbol{\beta})\partial_{ih}\ell(\boldsymbol{\beta})$, where $r_{ih}(\boldsymbol{\beta}) = y_{ih} - \pi_{ih}(\boldsymbol{\beta})$, which is an element of $T_{\boldsymbol{\beta}}\mathcal{S}$ and which measures the difference between a model in $\mathcal{M}$ and the observed data.

Even if our differential geometric framework is based on the assumption that we have k untied failure times, it can be extended to cases where tied events occur, such as the approach proposed in Kalbfleisch and Prentice (2002), in Cox (1972), in Breslow (1975) or in Efron (1977) for the Cox regression model. Here we will use the correction for tied events proposed in Cox (1972). Suppose we have d_i subjects failing at time t_i. By $\mathbf{i} = \{i_1 \ldots, i_{d_i}\}$ we denote the set of indices of the subjects falling at t_i and by $\mathcal{R}(t_i; d_i)$ the set of all possible subsets of d_i indices chosen from $\mathcal{R}(t_i)$ without replacement. With a little abuse of notation, the new multinomial random variable is denoted as $\mathbf{Y}_i = (Y_{i\mathbf{h}})_{\mathbf{h}\in\mathcal{R}(t_i;d_i)}$ and by $\boldsymbol{\pi}_i = (\pi_{i\mathbf{h}})_{\mathbf{h}\in\mathcal{R}(t_i;d_i)}$ the corresponding cell probabilities; under this setting the new ambient space is the set $\mathcal{S} = \left\{\prod_{i\in\mathcal{D}}\prod_{\mathbf{h}\in\mathcal{R}(t_i;d_i)} \pi_{i\mathbf{h}}^{y_{i\mathbf{h}}} : (\boldsymbol{\pi}_i)_{i\in\mathcal{D}} \in \bigotimes_{i\in\mathcal{D}} \Pi_i\right\}$. Finally, to complete our adjustment it is sufficient to change model (3.1) with a new model entailing the required adjusted partial likelihood. To this end, by setting

$$E_{\boldsymbol{\beta}}(Y_{i\mathbf{h}}) = \pi_{i\mathbf{h}}(\boldsymbol{\beta}) = \frac{\exp\{\boldsymbol{\beta}^\top \mathbf{s}_{\mathbf{h}}(t_i)\}}{\sum_{\mathbf{j}\in\mathcal{R}(t_i;d_i)}\exp\{\boldsymbol{\beta}^\top \mathbf{s}_{\mathbf{j}}(t_i)\}},$$

where $s_{\mathbf{j}}(t_i) = \sum_{l\in\mathbf{j}} \mathbf{x}_l(t_i)$, it is easy to see that the likelihood function associated with $\mathcal{M}$ is equal to the partial likelihood of the Cox regression model with the correction proposed in Cox (1972) when we assume that $y_{i\mathbf{h}}$ is equal to one if the set $\mathbf{h}$ is equal to $\mathbf{i}$ and zero otherwise. The same approach can also be used to handle the other corrections proposed in literature.

3.2. *dgLARS method for relative risk regression models*

dgLARS is a sequential method developed for constructing a sparse path of solutions indexed by a positive parameter γ and theoretically founded on the following characterization of the mth signed Rao score test statistic, i.e.:

$$r_m^u(\boldsymbol{\beta}) = I_{mm}^{-1/2}(\boldsymbol{\beta})\partial_m\ell(\boldsymbol{\beta}) = \cos\{\rho_m(\boldsymbol{\beta})\}\|r_{\boldsymbol{\beta}}\|_{\boldsymbol{\beta}}, \tag{3.3}$$

where $\|r_{\boldsymbol{\beta}}\|_{\boldsymbol{\beta}}^2 = \sum_{i\in\mathcal{D}}\sum_{h,k\in\mathcal{R}(t_i)} E_{\boldsymbol{\beta}}\{\partial_{ih}\ell(\boldsymbol{\beta})\partial_{ik}\ell(\boldsymbol{\beta})\}r_{ih}(\boldsymbol{\beta})r_{ik}(\boldsymbol{\beta})$ and $I_{mm}(\boldsymbol{\beta})$ is the Fisher information for β_m. The quantity $\rho_m(\boldsymbol{\beta})$ is a generalization of the Euclidean notion of angle between the mth column of the design matrix and the residual vector $\mathbf{r}(\boldsymbol{\beta}) = (r_{ih}(\boldsymbol{\beta}))_{i\in\mathcal{D},h\in\mathcal{R}(t_i)}$. Characterization (3.3) gives us a natural way to generalize the equiangularity condition of Efron *and others* (2004): two given predictors, say the mth and nth, satisfy the generalized 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 $\gamma^{(1)} \geq \cdots \geq \gamma^{(K)} \geq 0$, such that for any fixed γ between $\gamma^{(k+1)}$ and $\gamma^{(k)}$ the sub vector of the nonzero dgLARS estimates, denoted as $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma) = (\hat{\beta}_m(\gamma))_{m\in\hat{\mathcal{A}}}$, satisfies the following conditions:

$$r_m^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma)\} = s_m\gamma, \quad m \in \hat{\mathcal{A}} \tag{3.4}$$

$$|r_n^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma)\}| < \gamma, \quad n \notin \hat{\mathcal{A}} \tag{3.5}$$

where $s_m = \text{sign}\{\hat{\beta}_m(\gamma)\}$ and $\hat{\mathcal{A}} = \{m : \hat{\beta}_m(\gamma) \neq 0\}$, 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 occur:

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

$$|r_n^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}| = |r_m^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}| = \gamma^{(k)}, \tag{3.6}$$

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

$$\text{sign}[r_m^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma^{(k)})\}] \neq \text{sign}\{\hat{\beta}_m(\gamma^{(k)})\}, \tag{3.7}$$

   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 $\hat{\mathcal{A}} = \{a_1\}$, $\gamma^{(1)}$ 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 (3.6), then a_2 is included in $\hat{\mathcal{A}}$ and the new segment of the dgLARS curve is implicitly defined by the system with nonlinear equations:

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

where $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma) = (\hat{\beta}_{a_1}(\gamma), \hat{\beta}_{a_2}(\gamma))^\top$. 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 (3.6) occurs a new index is included in $\hat{\mathcal{A}}$ otherwise condition (3.7) occurs and an index is removed from $\hat{\mathcal{A}}$. 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 some fixed value that can be zero, if the sample size is large enough, or some positive value if we are working in a high-dimensional setting, i.e., the number of predictors is larger than the sample size. Table 1 reports the pseudocode of the developed algorithm to compute the dgLARS curve for a relative regression model. From a computational point of view, the entire dgLARS curve can be computed using the predictor–corrector algorithm proposed in Augugliaro *and others* (2013); for more details about this algorithm the interested reader is referred to Augugliaro *and others* (2014, 2016) or Pazira *and others* (2018). The latter extend the dgLARS method to GLM based on the exponential dispersion family proposing an improved predictor–corrector algorithm.

3.3. *Tuning parameter selection: derivation of the Generalized Information Criterion*

In the previous sections, we have seen how to extend the dgLARS method to relative regression models and how to construct the corresponding path of solutions. In this section, we address the problem of how to select the optimal point of such curve; in other words, the problem of finding the optimal γ-value. The behavior of any penalized estimator, such as Lasso or SCAD, is closely related to the way of selecting

Table 1. *Pseudocode of the dgLARS algorithm for a relative risk regression model*

Step	Description
0.	Let $r_m^u(\boldsymbol{\beta})$ be the Rao score statistic associated with the partial likelihood.
1.	Let $\gamma^{(1)} = \max_m \lvert r_m^u(\mathbf{0})\rvert$ and initialize the active set $\hat{\mathcal{A}} = \arg\max_m \lvert r_m^u(\mathbf{0})\rvert$
2.	Repeat the following steps
3.	Trace the segment of the dgLARS curve reducing γ and solving the system $r_m^u\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma)\} - s_m\gamma = 0, \quad m \in \hat{\mathcal{A}}$
4.	until γ is equal to the next transition point
5.	If condition (3.6) is met then include the new index in $\hat{\mathcal{A}}$
6.	Else (condition (3.7) is met) remove the index from $\hat{\mathcal{A}}$
7.	Until γ reaches some small positive value

the value of the tuning parameter because it controls the trade-off between bias and variance. Usually, the tuning parameter is selected using a suitable information criterion which can be written as:

$$\text{model fit} + C_n \times \text{ model complexity}, \tag{3.8}$$

where C_n is some positive sequence that depends only on the sample size. While minus two times the log-likelihood is commonly used as a measure of model fitting, there are many ways to measure model complexity. For example, this problem is studied for the Lasso estimator in Zou *and others* (2007).

In this article, we propose to select the optimal γ-value of the dgLARS method by using the Generalized Information Criterion (GIC) proposed in Konishi and Kitagawa (1996). For any fixed value of the parameter γ, dgLARS estimator can be seen as the Z-estimator implicitly defined by the system of estimating equations

$$\begin{aligned}\partial_m \ell_p\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma)\} - s_m\gamma I_{mm}^{1/2}\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma)\} &= \sum_{i\in\mathcal{D}}\left[\partial_m \ell_{i,p}\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma)\} - \gamma\frac{s_m I_{mm}^{1/2}\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma)\}}{|\mathcal{D}|}\right]\\ &= \sum_{i\in\mathcal{D}}\phi_{i,m}\{\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma)\}\\ &= 0, \quad m \in \hat{\mathcal{A}},\end{aligned}$$

where $\ell_p(\boldsymbol{\beta}) = \sum_{i\in\mathcal{D}} \ell_{i,p}(\boldsymbol{\beta})$ is the log-partial likelihood function. To the best of our knowledge, the approach developed in Konishi and Kitagawa (1996) is the only one providing a rigorous definition of model complexity for Z-estimators. Using definition (3.8), the GIC measure for the dgLARS estimator applied to the relative risk regression model is defined as

$$\text{GIC}(C_n) = -2\ell_p\{\hat{\boldsymbol{\beta}}(\gamma)\} + C_n \text{tr}\left[R^{-1}\{\hat{\boldsymbol{\beta}}(\gamma)\}Q\{\hat{\boldsymbol{\beta}}(\gamma)\}\right], \tag{3.9}$$

where $R\{\hat{\boldsymbol{\beta}}(\gamma)\}$ and $Q\{\hat{\boldsymbol{\beta}}(\gamma)\}$ are $|\hat{\mathcal{A}}| \times |\hat{\mathcal{A}}|$ matrices with generic elements

$$R_{mn}\{\hat{\boldsymbol{\beta}}(\gamma)\} = -\frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\frac{\partial\phi_{i,m}\{\hat{\boldsymbol{\beta}}(\gamma)\}}{\partial\beta_n} \quad \text{and} \quad Q_{mn}\{\hat{\boldsymbol{\beta}}(\gamma)\} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\phi_{i,m}\{\hat{\boldsymbol{\beta}}(\gamma)\}\partial_n \ell_{i,p}\{\hat{\boldsymbol{\beta}}(\gamma)\},$$

respectively.

For the case of the Cox regression model, i.e., $\psi\{\mathbf{x}(t);\boldsymbol{\beta}\} = \exp\{\boldsymbol{\beta}^\top \mathbf{x}(t)\}$, the log-partial likelihood function is equal to

$$\ell_p(\boldsymbol{\beta}) = \sum_{i\in\mathcal{D}} \left[\boldsymbol{\beta}^\top \mathbf{x}_i(t_i) - \log \sum_{j\in\mathcal{R}(t_i)} \exp\{\boldsymbol{\beta}^\top \mathbf{x}_j(t_i)\} \right]. \tag{3.10}$$

As shown in Cox (1972), we know that

$$\partial_m \ell_p(\boldsymbol{\beta}) = \sum_{i\in\mathcal{D}} \{x_{im}(t_i) - A_{i,m}(\boldsymbol{\beta})\}, \tag{3.11}$$

where $A_{i,m}(\boldsymbol{\beta}) = \sum_{j\in\mathcal{R}(t_i)} \exp\{\boldsymbol{\beta}^\top \mathbf{x}_j(t_i)\} x_{im}(t_i) / \sum_{j\in\mathcal{R}(t_i)} \exp\{\boldsymbol{\beta}^\top \mathbf{x}_j(t_i)\}$, and the mth element of the Fisher information matrix is equal to

$$I_{mn}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell_p(\boldsymbol{\beta})}{\partial\beta_m \partial\beta_n} = \sum_{i\in\mathcal{D}} \frac{\partial A_{i,m}(\boldsymbol{\beta})}{\partial\beta_n} = \sum_{i\in\mathcal{D}} \{C_{i,mn}(\boldsymbol{\beta}) - A_{i,m}(\boldsymbol{\beta}) A_{i,n}(\boldsymbol{\beta})\}, \tag{3.12}$$

where $C_{i,mn}(\boldsymbol{\beta}) = \sum_{j\in\mathcal{R}(t_i)} \exp\{\boldsymbol{\beta}^\top \mathbf{x}_j(t_i)\} x_{im}(t_i) x_{in}(t_i) / \sum_{j\in\mathcal{R}(t_i)} \exp\{\boldsymbol{\beta}^\top \mathbf{x}_j(t_i)\}$. Using (3.11) and (3.12), after straightforward algebra we have that

$$R_{mn}(\boldsymbol{\beta}) = \frac{1}{|\mathcal{D}|} \left\{ I_{mn}(\boldsymbol{\beta}) + \frac{s_m \gamma}{2 I_{mm}^{1/2}(\boldsymbol{\beta})} \frac{\partial I_{mm}(\boldsymbol{\beta})}{\partial\beta_n} \right\}, \tag{3.13}$$

where

$$\frac{\partial I_{mm}(\boldsymbol{\beta})}{\partial\beta_n} = \sum_{i\in\mathcal{D}} \left[C_{i,mmn}(\boldsymbol{\beta}) - C_{i,mm}(\boldsymbol{\beta}) A_{i,n}(\boldsymbol{\beta}) - 2 A_{i,m}(\boldsymbol{\beta}) \frac{\partial A_{i,m}(\boldsymbol{\beta})}{\partial\beta_n} \right],$$

and $C_{i,mmn}(\boldsymbol{\beta}) = \sum_{j\in\mathcal{R}(t_i)} \exp\{\boldsymbol{\beta}^\top \mathbf{x}_j(t_i)\} x_{im}^2(t_i) x_{in}(t_i) / \sum_{j\in\mathcal{R}(t_i)} \exp\{\boldsymbol{\beta}^\top \mathbf{x}_j(t_i)\}$. Finally,

$$Q_{mn}(\boldsymbol{\beta}) = \frac{1}{|\mathcal{D}|} \left\{ \sum_{i\in\mathcal{D}} \partial_m \ell_{i,p}(\boldsymbol{\beta}) \partial_n \ell_{i,p}(\boldsymbol{\beta}) - \frac{s_m \gamma I_{mm}^{1/2}(\boldsymbol{\beta}) \partial_n \ell_p(\boldsymbol{\beta})}{|\mathcal{D}|} \right\}. \tag{3.14}$$

The information criterion (3.9) is obtained evaluating the expressions (3.10), (3.13), and (3.14) at the point $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}(\gamma)$.

4. SIMULATION STUDY

In this section, we compare the method introduced in Section 3.2 with three popular algorithms: the coordinate descent method developed by Simon *and others* (2011), named CoxNet, the predictor–corrector developed by Park and Hastie (2007), named CoxPath, and the gradient ascent algorithm proposed by Goeman (2010), named CoxPen. These algorithms are implemented in the `R` packages `glmnet`, `glmpath`, and `penalized`, respectively. 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 referred to as the *dgCox* model.

4.1. *Comparison with other methods: the scale invariance*

Let $\ell_p(\boldsymbol{\beta})$ be the partial log-likelihood function, then the Lasso estimator is defined as solution of the problem

$$\max_{\boldsymbol{\beta}} \ell_p(\boldsymbol{\beta}) - \gamma \sum_{m=1}^{p} |\beta_m|, \tag{4.1}$$

where γ is a non-negative tuning parameter used to control the amount of sparsity in the resulting estimates; when γ is large some elements of the resulting Lasso estimates will be exactly equal to zero whereas when γ goes to zero the sparsity is reduced since more predictors will be included in the estimated model.

To better understand how the scale of the predictors influences the behavior of the Lasso estimator let $x_{im} = c_m z_{im}$, with $\sum_{i=1}^{n} z_{im} = 0$ and $\sum_{i=1}^{n} z_{im}^2 = 1$, then problem (4.1) is equivalent to the following reparameterized Lasso problem

$$\max_{\boldsymbol{\zeta}} \ell_p(\boldsymbol{\zeta}) - \sum_{m=1}^{p} \gamma_m |\zeta_m|, \tag{4.2}$$

where $\zeta_m = c_m \beta_m$ and $\gamma_m = \gamma / c_m$ is a tuning parameter specific for the mth regression coefficient. Problem (4.2) reveals that the scale factor of each predictor, i.e., the quantity c_m, influences the behavior of the Lasso estimator by changing the scale of the tuning parameter. For example, an increase of the factor c_m implies a reduction of the parameter γ_m consequently, with high probability, the mth predictor will included in the estimated model even if it does not influence the true relative risk function. The previous example shows that the ability of the Lasso estimator to select a set of predictors is not invariant under scale transformation of the predictors. dgLARS implicitly overcomes this theoretical limitation by using the Rao score test statistic and characterization (3.3) for variable selection. As consequence of the elementary properties of the Rao score test statistic, the variable selection property of the dgLARS method is scale invariant. For more details the reader is referred to Augugliaro *and others* (2016).

To study the practical effect, we perform a simulation study involving a Cox regression model where the survival time t_i is drawn from an exponential distribution with parameter $\lambda_i = \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)$ and $\mathbf{x}_i$ is sampled from a p-variate normal distribution $\mathrm{N}(\mathbf{0}, I)$, where I denotes the identity matrix. The censorship is randomly assigned to the survival times with probability $\pi = 0.5$. In our study, the sample size n is fixed to 100, p is fixed to 10 and β_m is equal to 0.5, for $m = 1, 2, 3$; the remaining 7 regression coefficients are 0.

To evaluate how the variable selection behavior is related to the scale of the predictors, we rescale in each simulation run the predictors with no effect to have Euclidean norm equal to k. In our study k varies from 1 to 4. Then we compute the path associated to dgCox, CoxNet, CoxPath, and CoxPen methods, respectively. For each point of a given path, denoted as $\hat{\boldsymbol{\beta}}(\gamma)$, we compute the False Positive Rate [FPR(γ)], i.e., the ratio between the number of false predictors selected by $\hat{\boldsymbol{\beta}}(\gamma)$ and the total number of false predictors, and the True Positive Rate [TPR(γ)], i.e., the ratio between the number of true predictors selected by $\hat{\boldsymbol{\beta}}(\gamma)$ and the total number of true predictors. These quantities are used to compute the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC). Finally, the four average AUCs are computed and studied as function of the Euclidean norm of the predictors with regression coefficient equal to zero. Figure 1 clearly confirms what we previously discussed, i.e., the variable selection behavior of the dgLARS method is scale invariant while the ability of the Lasso estimator to select the true model decreases as the k is increasing.

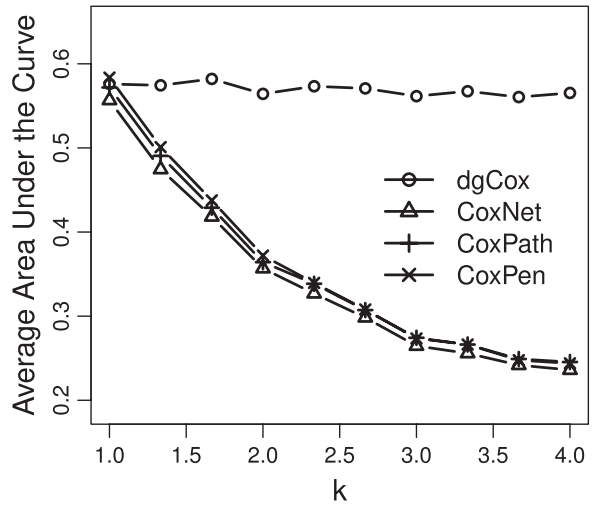

Fig. 1. Average area under the ROC curves seen as function of the Euclidean norm of the predictors with regression coefficient equal to zero.

4.2. *Global comparison with other path-estimation methods*

We simulated 100 datasets from a Cox regression model where the survival times t_i ($i = 1, \ldots, n$) follow an exponential distributions with parameter $\lambda_i = \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)$, and $\mathbf{x}_i$ 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 $\rho \in \{0.3, 0.9\}$. The censorship is randomly assigned to the survival times with probability $\pi \in \{0.2, 0.4\}$. The number of predictors is equal to 100 and the sample size is equal to 50 and 150. The first value is used to evaluate the behavior of the methods is a high-dimensional setting. Finally, we set $\beta_m = 0.2$ for $m = 1, \ldots, s$, where $s \in \{5, 10\}$; the remaining parameters are set equal to zero.

To remove the effects coming from the information measure used to select the optimal point of each path of solutions, we evaluated the global behavior of the paths by considering the ROC curves, which were computed as described in Section 4.1. For the sake of brevity, in Fig. 2 we show only the results for the simulation study with sample size equal to 50; the complete list of figures is reported in the supplementary material available at *Biostatistics* online. In this figure, we compare the ROC curves of our method with the three implementations of the lasso regularized Cox proportional hazards regression across six scenarios. In scenarios where $\rho = 0.3$, CoxNet, CoxPath, and CoxPen exhibit a similar performance, having overlapping curves for both levels of censoring, whereas dgCox method appears to be consistently better with the largest AUC. A similar performance of the methods has been also observed for the other combinations of the ρ and π values. In scenarios where the correlation among neighboring predictors is high, i.e., $\rho = 0.9$, the dgCox method is clearly the superior approach for all levels of censoring. As shown in the figures reported in the supplementary material available at *Biostatistics* online, a similar result is obtained when we increase the sample size.

4.3. *Tuning parameter selection comparisons*

As seen in Section 3.3, the behavior of any penalized method is closely related to the information criterion used to select the optimal point of the path of solutions. The simulation study reported in this section is intended to examine the finite sample performance of a number of model information criteria applied to the

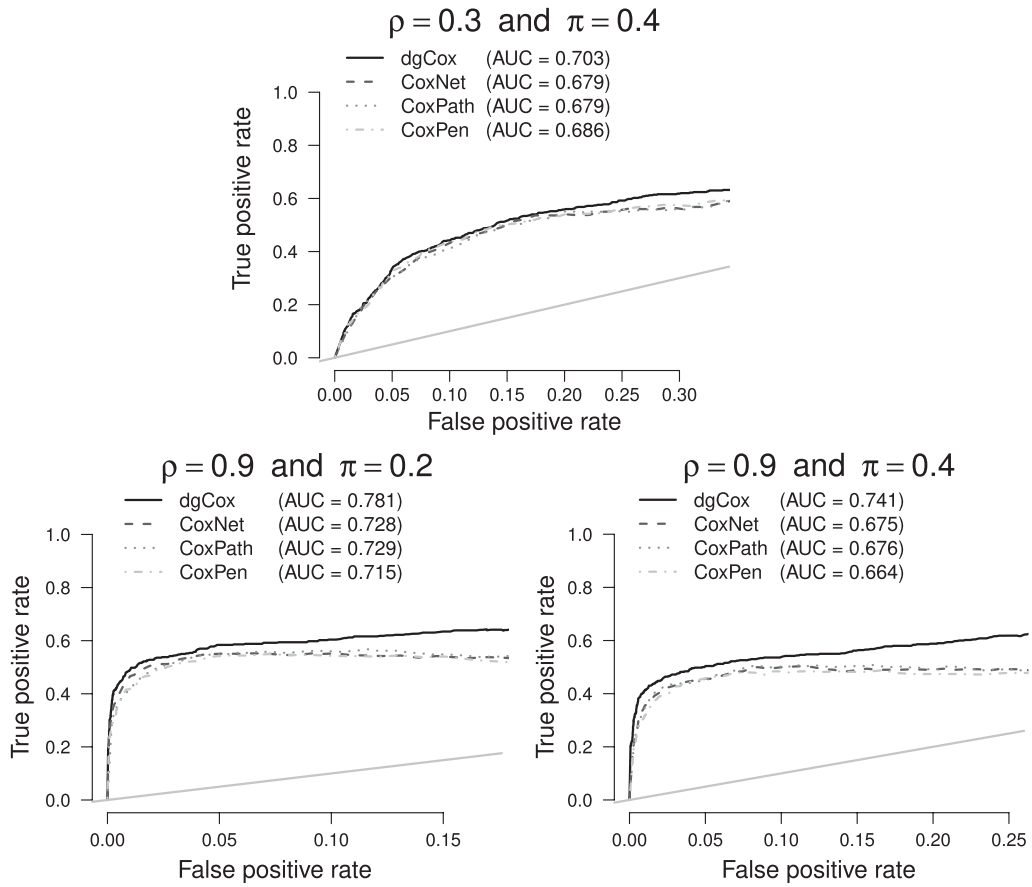

Fig. 2. Results from the simulation study with $s = 5$ and sample size equal to 50; for each scenario, we show the averaged ROC curve for dgCox, CoxNet, CoxPath, and CoxPen algorithm. The average AUC is also reported. The 45° diagonal is also included in the plots.

dgCox model. The measures that we considered in our study are the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the GIC proposed in Fan and Tang (2013) (FAN13), which uses $C_n = \log(\log|\mathcal{D}|)\log p$ and complexity term $|\hat{\mathcal{A}}(\gamma)|$, and two possible versions of the GIC measure (3.9), with $C_n = 2$, as originally proposed, and $C_n = \log|\mathcal{D}|$, imitating the BIC. All the considered criteria are based on using the partial log-likelihood as measure of model fit.

The simulation study in this section follows a similar data generation mechanism as discussed in Section 4.2; we fix the censoring probability $\pi = 0.2$, the sample sizes $n \in \{50, 200\}$ and $p \in \{50, 100, 1000\}$. This scenario also covers the high-dimensional setting. The level of sparsity in the true model varies: for the first $d = \{2, 8, 32\}$ predictors we fix the coefficients to 0.5, whereas the remaining coefficients are set to zero. The same correlation structure Σ as in Section 4.2 is considered with $\rho = 0.9$. For each scenario, we simulate 500 data sets and let the dgCox method computes the entire path of solutions. Then, we use the considered information criteria to select the optimal γ-value.

For the sake of brevity, in Table 2 we report only the results of the simulation study with sample size equal to $n = 50$; the remaining results are reported in the supplementary material available at *Biostatistics* online. To evaluate the behavior of the considered criteria we use the median number of variables included

Table 2. *Results from the simulation studies when $n = 50$; for each scenario we report the median number of variables included in the final model (Size), the mean of the false positive rate (FPR), the false discovery rate (FDR), the false negative rate (FNR), and F1-score (F1). Standard errors are in parentheses. Bold values identify the best information criterion for each scenario*

p	d	Criterion	Size	FPR	FDR	FNR	F1
50	2	AIC	4 (0.331)	0.071 (0.007)	0.539 (0.030)	**0.220** (0.027)	0.523 (0.024)
		BIC	2 (0.159)	0.023 (0.003)	0.310 (0.029)	0.295 (0.029)	0.635 (0.023)
		FAN13	2 (0.105)	**0.011** (0.002)	**0.216** (0.028)	0.365 (0.030)	**0.648** (0.025)
		GIC(2)	4 (0.302)	0.061 (0.006)	0.507 (0.030)	0.230 (0.027)	0.541 (0.023)
		GIC($\log \lvert\mathcal{D}\rvert$)	2 (0.166)	0.061 (0.006)	0.507 (0.030)	0.230 (0.027)	0.541 (0.023)
	8	AIC	9 (0.288)	0.080 (0.006)	0.313 (0.017)	**0.236** (0.013)	0.708 (0.011)
		BIC	7 (0.185)	0.032 (0.003)	0.166 (0.014)	0.289 (0.014)	0.756 (0.011)
		FAN13	6 (0.162)	**0.019** (0.002)	**0.108** (0.012)	0.319 (0.015)	**0.760** (0.011)
		GIC(2)	9 (0.284)	0.081 (0.006)	0.321 (0.016)	**0.236** (0.013)	0.705 (0.012)
		GIC($\log \lvert\mathcal{D}\rvert$)	7 (0.175)	0.081 (0.006)	0.321 (0.016)	**0.236** (0.013)	0.705 (0.012)
	32	AIC	16 (0.224)	0.040 (0.004)	0.044 (0.004)	0.522 (0.006)	0.635 (0.006)
		BIC	14 (0.224)	0.033 (0.003)	0.040 (0.004)	0.554 (0.007)	0.606 (0.007)
		FAN13	13 (0.322)	**0.026** (0.003)	**0.036** (0.005)	0.610 (0.010)	0.548 (0.011)
		GIC(2)	16 (0.222)	0.041 (0.004)	0.045 (0.004)	**0.519** (0.006)	**0.637** (0.006)
		GIC($\log \lvert\mathcal{D}\rvert$)	15 (0.237)	0.041 (0.004)	0.045 (0.004)	**0.519** (0.006)	**0.637** (0.006)
1000	2	AIC	8 (0.519)	0.007 (0.001)	0.723 (0.026)	**0.310** (0.026)	0.344 (0.023)
		BIC	2 (0.211)	**0.001** (0.000)	**0.359** (0.035)	0.405 (0.026)	**0.551** (0.026)
		FAN13	1 (0.081)	0.000 (0.000)	0.152 (0.029)	0.640 (0.031)	0.426 (0.034)
		GIC(2)	7 (0.516)	0.007 (0.001)	0.702 (0.027)	**0.310** (0.026)	0.362 (0.024)
		GIC($\log \lvert\mathcal{D}\rvert$)	2 (0.215)	0.007 (0.001)	0.702 (0.027)	**0.310** (0.026)	0.362 (0.024)
	8	AIC	9 (0.377)	0.004 (0.000)	0.388 (0.023)	**0.380** (0.014)	0.584 (0.011)
		BIC	5 (0.229)	0.001 (0.000)	0.151 (0.018)	0.429 (0.013)	**0.662** (0.011)
		FAN13	4 (0.124)	**0.000** (0.000)	**0.002** (0.002)	0.545 (0.015)	0.608 (0.016)
		GIC(2)	10 (0.395)	0.005 (0.000)	0.401 (0.024)	0.381 (0.014)	0.576 (0.012)
		GIC($\log \lvert\mathcal{D}\rvert$)	6 (0.258)	0.005 (0.000)	0.401 (0.024)	0.381 (0.014)	0.576 (0.012)
	32	AIC	14 (0.212)	0.001 (0.000)	0.056 (0.007)	0.572 (0.006)	0.586 (0.006)
		BIC	14 (0.229)	0.001 (0.000)	0.047 (0.006)	0.593 (0.007)	0.567 (0.007)
		FAN13	1 (0.399)	**0.000** (0.000)	**0.007** (0.003)	0.898 (0.012)	0.163 (0.019)
		GIC(2)	15 (0.219)	0.001 (0.000)	0.061 (0.007)	**0.570** (0.006)	**0.587** (0.006)
		GIC($\log \lvert\mathcal{D}\rvert$)	14 (0.231)	0.001 (0.000)	0.061 (0.007)	**0.570** (0.006)	**0.587** (0.006)

in the final model (Size), the average false positive rate (FPR), the false discovery rate (FDR), the false negative rate (FNR), and F1-score (F1) to investigate the performance of the model selection criteria in identifying the true model. The results in Table 2 show that there is a clear trade-off between FNR and FPR. Therefore, it can be more informative to compare a summary measure, such as the F1-score. Summarizing, we found that in very sparse contexts, i.e., the true number of effects is small ($d \ll p$), FAN13 performs well in terms of F1-score. The two GIC measures perform well when d increases, slightly beating FAN13. The traditional AIC and BIC criteria do not perform as well: the reason could be that the penalized partial likelihood setting violate the usual assumptions for these methods in that model fit should be measured as a maximum likelihood, not as a penalized partial likelihood.

5. Finding genetic signatures in cancer survival

In this section, we test the predictive power of proposed method in four recent studies. In particular, we focus on the identification of genes involved in the regulation of colon cancer (Loboda *and others*, 2011), prostate cancer (Ross *and others*, 2012), ovarian cancer (Gillet *and others*, 2012), and skin cancer (Jönsson *and others*, 2010). The set-up of the four studies was similar. In the patient a cancer was detected and treated. When treatment was complete a follow-up started. In all cases, the expression of several genes were measured in the affected tissue together with the survival times of the patients, which may be censored if the patients were alive when they left the study. Although other socio-economical variables, such us age, sex, etc. are available, our analysis only focuses on the impact of the gene expression levels on the patients' survival.

A table containing a brief description of the four datasets used in this section is reported in the supplementary materials available at *Biostatistics* online. In the four scenarios, the number of predictors p is larger than the number of patients n. The dimensionality is especially high in the cases of the colon and skin cancer where the expression of several thousands of genes were measured. In the prostate and ovarian cancers the number of genes is 162 and 306, which will also help us to study the performance of the dgLARS method when the number of variables is just a few orders of magnitude larger than the number of observations.

In genomics, it is common to assume that just a moderate number of genes affect the phenotype of interest. To identify such genes in this survival context, we estimate a Cox regression model using the dgLARS method described in Section 3. To this end, we randomly select a training sample that contains the 60% of the patients, and we save the remaining data to test the models. We calculate the paths of solutions in the four scenarios and we select the optimal number of components by means of the GIC($\log|\mathcal{D}|$) criterion derived in Section 3.3. For the colon, prostate, ovarian, and skin cancer studies we find gene profiles consisting of, respectively, 38, 24, 43, and 23 genes.

In order to illustrate the prediction performance of the dgLARS method, we classify the test patients into a low-risk group and a high-risk groups by splitting the test sample into two subsets of equal size according to the estimated individual predicted excess risk. The first two plots in Fig. 3 show the Kaplan–Maier survival curves estimates for the low- and high-risk groups together with the original training survival curve for the Colon and Ovarian studies, respectively. To test the groups separation we use the non-parametric Peto & Peto modification of the Gehan–Wilcoxon test (Peto and Peto, 1972). For all four studies, the differences between the low- and high-risk groups are significant at the traditional 0.05 significance level.

Alternatively, survival ROC curves (Heagerty *and others*, 2000) can be used to describe how well the selected model is predicting the order of survival of the patients in each of the studies. It can be interpreted as a traditional ROC curve with respect to predicting the survival of the patients. The final two plots in Fig. 3 show the survival ROC curves for the relatively small ovarian cancer study and the large colon cancer study. The results show that dgCox combined with GIC is better than other sparse survival methods in predicting the survival order. The results for the other two datasets are even more in favor of the dgCox method and are reported in the supplementary materials available at *Biostatistics* online. Both results suggest that the reported gene profiles are predictive for determining survival and it demonstrates the power of dgLARS as a tool in medical analysis for massive gene screening studies.

To gain some biological understanding of the process of cancer regulation we performed an enrichment analysis of the 21 genes that have been found to be relevant in the regulation of the skin cancer. For the sake of brevity the enrichment analysis is reported in the supplementary material available at *Biostatistics* online.

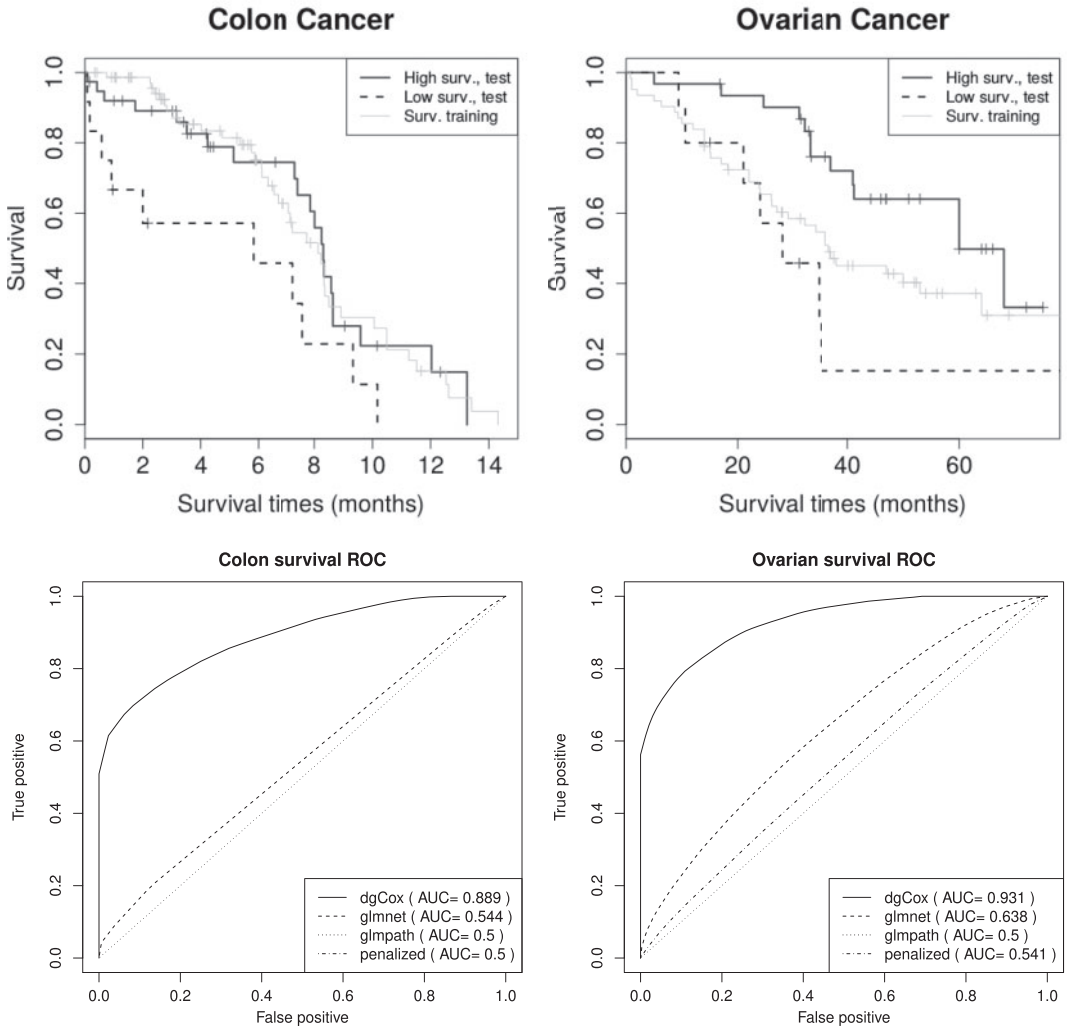

Fig. 3. Top plots: the Kaplan–Meier survival curves for colon and ovarian studies on *training* data together with the curves associated to the low- and high-risk groups in the *test* sample, showing dgCox's ability being able to detect low- and high-risk individuals. Bottom plots: survival ROC curves for the same two studies, demonstrating how dgCox combined with GIC has a better predictive performance than other sparse survival methods.

6. CONCLUSIONS

In this article, we have extended the dgLARS method to relative risk regression model using the relationship exiting between the partial likelihood function and a specific GLM. The advantage of this approach is that the estimates are invariant to arbitrary changes in the measurement scales of the predictors. Unlike SCAD or ℓ_1 sparse regression methods, no prior rescaling of the predictors is therefore needed. The proposed method can be used for a large class of survival models, the so called relative risk models. We have implementations for the Cox proportional hazards model and the excess relative risk model.

In this article, we have also proposed a new information criterion to select the optimal point in the path of solutions defined by applying dgLARS method to a relative risk regression model. As our method

involves shrinkage of the parameters, the issue of the underlying degrees of freedom of the sparse models is a complex one. For this reason, we used the approach developed in Konishi and Kitagawa (1996), which provides a rigorous definition of model complexity for Z-estimators. We showed that the proposed measure works well in a simulation study and is only beaten the method from Fan and Tang (2013) when the true underlying model is extremely sparse.

A software implementation of our method is available on github (https://github.com/LuigiAugugliaro/dgcox) and can deal with the classical $n > p$ setting, but also with the high-dimensional setting, such as, for example, a skin cancer study with $p = 30\,807$ predictors and $n = 54$ observations. We have considered four recent cancer survival studies, where we look for a genetic "survival signature." Due to the large number of predictors, the studies are unsuitable for traditional survival regression methods. Instead, the results we find go beyond univariate importance and by means of an enrichment study can be linked to potentially relevant biological explanations.

Supplementary materials

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors gratefully acknowledge the editor and two referees, whose comments have improved the quality of this work.

Conflict of Interest: None declared.

Funding

The authors have received funding for collaboration through the EU COST Action CA15109 (COSTNET).

References

Augugliaro, L., Mineo, A. M. and Wit, E. C. (2013). Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models. *Journal of the Royal Statistical Society Series B* **75**, 471–498.

Augugliaro, L., Mineo, A. M. and Wit, E. C. (2014). dglars: an R package to estimate sparse generalized linear models. *Journal of Statistical Software* **59**, 1–40.

Augugliaro, L., Mineo, A. M. and Wit, E. C. (2016). A differential geometric approach to generalized linear models with grouped predictors. *Biometrika* **103**, 563–577.

Breslow, N. E. (1975). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D. R. (1981). Discussion of paper by D. Oakse entitled "Survival times: aspects of partial likelihood". *International Statistical Review* **49**, 258.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Monographs on Statistics and Applied Probability. London: Chapman and Hall.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72**, 557–565.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.

FAN, J. AND LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

FAN, Y. AND TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B* **75**, 531–552.

GILLET, J. P., CALCAGNO, A. M., VARMA, S., DAVIDSON, B., ELSTRAND, M. B., GANAPATHI, R., KAMAT, A. A., SOOD, A. K., AMBUDKAR, S. V., SEIDEN, M. V. *and others*. (2012). Multidrug resistance-linked gene signature predicts overall survival of patients with primary ovarian serous carcinoma. *Clinical Cancer Research* **18**, 3197–3206.

GOEMAN, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* **52**, 70–84.

GUI, J. AND LI, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21**, 3001–3008.

HEAGERTY, P. J., LUMLEY, T. AND PEPE, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

JÖNSSON, G., BUSCH, C., KNAPPSKOG, S., GEISLER, J., MILETIC, H., RINGNÉR, LILLEHAUG, J. R., BORG, A. AND LØNNING, P. E. (2010). Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clinical Cancer Research* **16**, 3356–3367.

KALBFLEISCH, J. D. AND PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc.

KONISHI, S. AND KITAGAWA, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–890.

LOBODA, A., NEBOZHYN, M. V., WATTERS, J. W., BUSER, C. A., SHAW, P. M., Huang, P. S., L. Van't Veer, R. A. Tollenaar, Jackson, D. B., Agrawal, D. *and others*. (2011). EMT is the dominant program in human colon cancer. *BMC Medical Genomics*, **4**, 9.

MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman & Hall.

MOOLGAVKAR, S. H. AND VENZON, D. J. (1987). Confidence regions in curved exponential families: application to matched case-control and survival studies with general relative risk function. *The Annals of Statistics* **15**, 346–359.

OAKES, D. (1981). Survival times: aspects of partial likelihood. *International Statistical Review* **49**, 235–252.

PARK, M. Y. AND HASTIE, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B* **69**, 659–677.

PAZIRA, H., AUGUGLIARO, L. AND WIT, E. C. (2018). Extended differential geometric LARS for high-dimensional GLMs with general dispersion parameter. *Statistics and Computing* **28**, 753–774.

PETO, R. AND PETO, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society Series A* **135**, 185–207.

PRENTICE, R. L. AND BRESLOW, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65**, 153–158.

PRENTICE, R. L. AND MASON, M. W. (1996). On the application of linear relative risk regression models. *Biometrics* **42**, 109–120.

PRENTICE, R. L., YOSHIMOTO, Y. AND MASON, M. (1983). Relationship of cigarette smoking and radiation exposure to cancer mortality in Hiroshima and Nagasaki. *Journal of National Cancer Institute* **70**, 611–622.

RAO, C. R. (1949). On the distance between two populations. *Sankhyā* **9**, 246–248.

RIPPE, R. C. A., MEULMAN, J. J. AND EILERS, P. H. C. (2012). Visualization of genomic changes by segmented smoothing using an L_0 penalty. *PLoS One* **7**, e38230.

ROSS, R. W., GALSKY, M. D., SCHER, H. I., MAGIDSON, J., WASSMANN, K., LEE, G. S. M., KATZ, L., SUBUDHI, S. K., ANAND, A., FLEISHER, M., KANTOFF, P. W. *and others*. (2012). A whole-blood RNA transcript-based prognostic model in men with castration-resistant prostate cancer: a prospective study. *The Lancet Oncology* **13**, 1105–1113.

SIMON, N., FRIEDMAN, J. H., HASTIE, T. AND TIBSHIRANI, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* **39**, 1–13.

SOHN, I., KIM, J., JUNG, S. H. AND Park, C. (2009). Gradient lasso for Cox proportional hazards model. *Bioinformatics* **25**, 1775–1781.

THOMAS, D. C. (1977). Addendum to the paper by Liddell, McDonald, Thomas and Cunliffe. *Journal of the Royal Statistical Society Series A* **140**, 483–485.

THOMAS, D. C. (1981). General relative-risk models for survival time and matched case-control analysis. *Biometrics* **37**, 673–686.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in medicine* **16**, 385–395.

ZOU, H. AND HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 301–320.

ZOU, H., HASTIE, T. AND TIBSHIRANI, R. (2007). On the "degrees of freedom" of the LASSO. *The Annals of Statistics* **35**, 2173–2192.

[*Received May 13, 2017; revised September 20, 2018; accepted for publication September 24, 2018*]