DgCox: a differential geometric approach for high-dimensional Cox proportional hazard models

Ernst Wit[(1)], Luigi Augugliaro[(2)], Fentaw Abegaz[(1)], Javier Gonzalez[(3)]

(1) University of Groningen
Johann Bernoulli Institute, Nijenborgh 9, 9747 AG, Groningen, The Netherlands,
e.c.wit@rug.nl, f.abegaz.yazew@rug.nl.

(2) University of Palermo
Department of Statistical and Mathematical Sciences "S. Vianelli", Palermo, Italy,
luigi.augugliaro@unipa.it.

(3) University of Sheffield
Machine Learning Group, Glossop Road S10 2HQ, Sheffield, UK.
j.h.gonzalez@sheffield.ac.uk.

Keywords: Proportional hazard modelling, least angle regression, differential geometry, sparse inference.

Abstract. Many clinical and epidemiological studies rely on survival modelling to detect clinically relevant factors that affect various event histories. With the introduction of high-throughput technologies in the clinical and even large-scale epidemiological studies, the need for inference 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. This paper will introduce a principled sparse inference methodology for proportional hazards modelling, based on differential geometrical analyses of the high-dimensional likelihood surface.

1 Scientific background

In the study of the dependence of survival time T on covariates $x = (x_1, \ldots, x_p)$, the Cox proportional hazards model (Cox, 1972, 1975) has proved to be a major tool in many clinical and epidemiological applications. Especially when the number of explanatory variables are small, the standard model combined with regression-like structure learning hypothesis testing is adequate in most situations. However, when the number of features is large, the simple Cox proportional breaks down. However, in practice, not all the covariates may contribute to the prediction of survival outcomes: some components of coefficients may be zero in the true model.

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 (Tibshirani, 1997) and the smoothly clipped absolute deviation method (Fan and Li, 2002) 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 least angle regression approach (Efron et al., 2004) to the case of the Cox proportional hazards model. This means that we have to explore the geometry of the underlying likelihood in order to define efficient computational methods for the solution.

2 Differential geometric Cox proportional hazards model

The aim of this section is to extend the dgLARS method (Augugliaro et al., 2013) to the relative risk regression models described in section 2.1. 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 originally proposed in Efron et al. (2004). We first relate the partial likelihood with the likelihood function of a specific GLM. As originally observed in Thomas (1977), to solve this problem we shall use the identity that exists between the partial likelihood (3) 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 in matched case-control studies.

2.1 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 death 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

$$\lambda(t) = \frac{f(t)}{1 - \int_0^t f(s)ds}, \tag{1}$$

and specifies the instantaneous rate at which failures occur for subjects that are surviving at time t. Suppose that the hazard function (1) can depend on a p-dimensional vector of covariates which can depend on time and is denoted by $\mathbf{x}(t) = (x_1(t), \ldots, x_p(t))^T$. The relative risk regression models are based on the assumption that the vector $\mathbf{x}(t)$ influence the hazard function by the following relation

$$\lambda(t; \mathbf{x}) = \lambda_0(t)\psi(\mathbf{x}(t); \boldsymbol{\beta}), \tag{2}$$

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 hazard function at time t which is left unspecified. Finally, $\psi : \mathbb{R} \to \mathbb{R}$ is a fixed twice continuously 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. $\psi(\mathbf{0}; \boldsymbol{\beta}) = 0$. Model (2), originally proposed in Thomas (1981), extends the usual Cox regression model Cox (1972), which is obtained when $\psi(\mathbf{x}(t); \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^T\mathbf{x}(t))$.

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 by $\mathcal{D}$ 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 how have not failed and are still under observation just prior to time t, under the assumption of independent censoring, the inference about the $\boldsymbol{\beta}$ can be carried out by the following partial likelihood function

$$\mathcal{L}_p(\boldsymbol{\beta}) = \prod_{i \in \mathcal{D}} \frac{\psi(\mathbf{x}_i(t_i); \boldsymbol{\beta})}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i); \boldsymbol{\beta})}. \tag{3}$$

When the exponential relative risk function is used in model (2) and we work with fixed covariates, (3) is clearly equal to the original partial likelihood introduced in Cox (1972) and discussed in great detail in Cox (1975). The inferential aspects of the relative risk regression models (2) are studied in Prentice and Self (1983) where are extended the results given in Andersen and Gill (1982) for the Cox regression model.

2.2 Differential geometric description of the relative risk regression model

In order to define the generalized equiangularity condition for the relative risk regression model, it is useful to see the partial likelihood (3) as arising from a multinomial sampling 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 model space

$$\mathcal{S} = \left\{ \prod_{i\in\mathcal{D}} \prod_{h\in\mathcal{R}(t_i)} \pi_{ih}^{y_{ih}} : (\boldsymbol{\pi}_i)_{i\in\mathcal{D}} \in \bigotimes_{i\in\mathcal{D}} \Pi_i \right\}. \tag{4}$$

In the following of our differential geometric constructions, the set (4) will play the role of ambient space. We would like to stress that our differential geometric constructions are invariant to the chosen parameterization which means that ambient space $\mathcal{S}$ can be equivalently defined by the canonical parameter vector and this will not change our results. In this paper we prefer to use the mean value parameter vector to specify our differential geometrical tools because in this way the relationship with the partial likelihood (3) will be more transparent.

Consider the following model for the conditional expected value of the random variable Y_{ih}, i.e.

$$E_{\boldsymbol{\beta}}(Y_{ih}) = \pi_{ih}(\boldsymbol{\beta}) = \frac{\psi(\mathbf{x}_h(t_i);\boldsymbol{\beta})}{\sum_{j\in\mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i);\boldsymbol{\beta})}, \tag{5}$$

then our model space is the set

$$\mathcal{M} = \left\{ \prod_{i\in\mathcal{D}} \prod_{h\in\mathcal{R}(t_i)} \left(\frac{\psi(\mathbf{x}_h(t_i);\boldsymbol{\beta})}{\sum_{j\in\mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i);\boldsymbol{\beta})} \right)^{y_{ih}} : \boldsymbol{\beta} \in \mathcal{B} \right\}. \tag{6}$$

The partial likelihood (3) is formally equivalent to the likelihood function associated with the model space $\mathcal{M}$ if we assume 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 that 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 the standard regularity conditions, it is easy to see that $T_{\boldsymbol{\beta}}\mathcal{M}$ is the linear vector space of the random variables $v_{\boldsymbol{\beta}} = \sum_{m=1}^{p} v_m \partial_m \ell(\boldsymbol{\beta}) \in T_{\boldsymbol{\beta}}\mathcal{M}$ with zero expected value and finite variance, i.e.

$$E_{\boldsymbol{\beta}}(v_{\boldsymbol{\beta}}) = 0 \quad \text{and} \quad E_{\boldsymbol{\beta}}(v_{\boldsymbol{\beta}}^2) < \infty.$$

As a simple consequence of the chain rule we have the following identity for any tangent vector belonging to the tangent space $T_{\boldsymbol{\beta}}\mathcal{M}$,

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

which shows that $T_{\boldsymbol{\beta}}\mathcal{M}$ is a linear vector subspace of the tangent space $T_{\boldsymbol{\beta}}\mathcal{S}$ 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 $T_{\boldsymbol{\beta}}\mathcal{M}$, 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; Burbea and Rao, 1982), i.e

$$\langle v_{\boldsymbol{\beta}}; w_{\boldsymbol{\beta}} \rangle_{\boldsymbol{\beta}} = E_{\boldsymbol{\beta}}(v_{\boldsymbol{\beta}} \cdot w_{\boldsymbol{\beta}}) = \sum_{m,n=1}^{p} E_{\boldsymbol{\beta}} \left(\partial_m \ell(\boldsymbol{\beta}) \cdot \partial_n \ell(\boldsymbol{\beta}) \right) v_m w_m = \mathbf{v}^T I(\boldsymbol{\beta}) \mathbf{w}, \quad (7)$$

where $\mathbf{v} = (v_1, \ldots, v_p)$, $\mathbf{w} = (w_1, \ldots, w_p)$ 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 (7) 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 (Prentice and Self, 1983).

As we shall see in the next section, the dgLARS estimator is based on a differential geometric characterization of the Rao score test statistic which is obtained considering the inner product between the bases of the tangent space $T_{\boldsymbol{\beta}}\mathcal{M}$ and the following tangent residual vector

$$r_{\boldsymbol{\beta}} = \sum_{i \in \mathcal{D}} \sum_{h \in \mathcal{R}(t_i)} r_{ih}(\boldsymbol{\beta}) \partial_{ih} \ell(\boldsymbol{\beta}), \quad (8)$$

where $r_{ih}(\boldsymbol{\beta}) = y_{ih} - \pi_{ih}(\boldsymbol{\beta})$. Since in the definition (8), the observed response vector $\mathbf{y}$ is considered fixed, it is easy to see that $r_{\boldsymbol{\beta}} \in T_{\boldsymbol{\beta}}\mathcal{S}$.

2.3 DgCox: dgLARS method for the relative risk regression model

The dgLARS method is a sequential method developed to estimate a sparse solution curve embedded in the in the parameter space $\mathcal{B}$. To explore the sparse structure of a relative risk regression model, we can use the following differential geometric characterization characterization of the mth element of the score vector, i.e.

$$\partial_m \ell(\boldsymbol{\beta}) = \langle \partial_m \ell(\boldsymbol{\beta}); r_{\boldsymbol{\beta}} \rangle_{\boldsymbol{\beta}} = \cos(\rho_m(\boldsymbol{\beta})) \cdot I_{mm}^{1/2}(\boldsymbol{\beta}) \cdot \|r_{\boldsymbol{\beta}}\|_{\boldsymbol{\beta}}, \quad (9)$$

where $I_{mm}(\boldsymbol{\beta})$ is the Fisher information for β_m, $\|r_{\boldsymbol{\beta}}\|^2_{\boldsymbol{\beta}}$ is equal to

$$\begin{aligned} E_{\boldsymbol{\beta}}(r^2(\boldsymbol{\beta})) &= \sum_{i,j \in \mathcal{D}} \sum_{h \in \mathcal{R}(t_i)} \sum_{k \in \mathcal{R}(t_j)} E_{\boldsymbol{\beta}}(\partial_{ih}\ell(\boldsymbol{\beta}) \cdot \partial_{jk}\ell(\boldsymbol{\beta})) \, r_{ih}(\boldsymbol{\beta}) r_{jk}(\boldsymbol{\beta}) = \\ &= \sum_{i \in \mathcal{D}} \sum_{h,k \in \mathcal{R}(t_i)} E_{\boldsymbol{\beta}}(\partial_{ih}\ell(\boldsymbol{\beta}) \cdot \partial_{ik}\ell(\boldsymbol{\beta})) \, r_{ih}(\boldsymbol{\beta}) r_{ik}(\boldsymbol{\beta}) = \\ &= \sum_{i \in \mathcal{D}} \sum_{h,k \in \mathcal{R}(t_i)} \frac{r_{ih}(\boldsymbol{\beta}) r_{ik}(\boldsymbol{\beta})}{\pi_{ih}(\boldsymbol{\beta}) 1_{\{h=k\}} - \pi_{ih}(\boldsymbol{\beta}) \pi_{ik}(\boldsymbol{\beta})} \end{aligned}$$

and $\rho_m(\boldsymbol{\beta})$ is a generalization of the Euclidean notion of angle between the mth column of the design matrix and the residual vector $\mathbf{r}(\boldsymbol{\beta}) = (r_{ih}(\boldsymbol{\beta}))_{i \in \mathcal{D}, h \in \mathcal{R}(t_i)}$. Using (9) one can see that the signed Rao's score test statistic can be geometrically characterized as follows:

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

then we shall say that two given predictors, say m and n, 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 active set. Formally, for a given value of the tuning parameter $\gamma \in \mathbb{R}^+$

Table 1: Overview of the dgCox method to compute the solution curve

Step	*DgCox algorithm*
1	start with the intercept only model
2	repeat
3	increase the parameters of the active variables keeping the angles between their scores and residual tangent vector the same
4	if the angle of a not-included variable is the same as the ones currently in the model include that variable in the active set
5	until a stopping rule is met

the corresponding active set is denoted by $\hat{\mathcal{A}}(\gamma)$ and the dgLARS estimator, denoted by $\hat{\boldsymbol{\beta}}(\gamma)$, is such that the following conditions are satisfied:

$$\forall\, m \in \hat{\mathcal{A}}(\gamma) \quad \rightarrow \quad \left| r_m^u(\hat{\boldsymbol{\beta}}(\gamma)) \right| = \gamma, \tag{10}$$

$$r_m^u(\hat{\boldsymbol{\beta}}(\gamma)) = s_m \gamma, \tag{11}$$

$$\forall\, n \notin \hat{\mathcal{A}}(\gamma) \quad \rightarrow \quad |r_n^u(\hat{\boldsymbol{\beta}}(\gamma))| < \gamma, \tag{12}$$

where $s_m = \text{sign}(\hat{\beta}_m(\gamma))$.

Using the differential geometrical structure of a relative risk regression model explained in section 2.2 and the previous conditions, it is possible to use the dgLARS method to explore the sparse structure of a relative risk regression model. Formally, the dgLARS method computes a finite sequence of transition points, say $0 \leq \gamma^{(K)} \leq \ldots \leq \gamma^{(2)} \leq \gamma^{(1)}$, such that for each $\gamma^{(k)}$ one of the following two conditions can occur:

(i) $\exists\, n \notin \hat{\mathcal{A}}(\gamma^{(k-1)})$ such that $\left| r_n^u(\hat{\boldsymbol{\beta}}(\gamma^{(k)})) \right| = \gamma$, and therefore $\hat{\mathcal{A}}(\gamma^{(k)}) = \hat{\mathcal{A}}(\gamma^{(k-1)}) \cup \{n\}$;

(ii) $\exists\, m \in \hat{\mathcal{A}}(\gamma^{(k-1)})$ such that $\text{sign}(r_m^u(\hat{\boldsymbol{\beta}}(\gamma^{(k)}))) \neq \text{sign}(\hat{\beta}_m(\gamma^{(k)}))$, and therefore $\hat{\mathcal{A}}(\gamma^{(k)}) = \hat{\mathcal{A}}(\gamma^{(k-1)}) \setminus \{m\}$,

which means that a new predictor is included in the active set when the generalized equiangularity condition is satisfied, namely condition (i), or an active predictor is removed from the active set if the sign of the corresponding signed Rao's score test statistic is not in agreement with the sign of the estimated coefficient, i.e. condition (ii). Table 1 shows in a concise form the outline of the dgCox algorithm.

3 Simulation study

In this section we compare the dgCox model with there popular algorithms for sparse Cox regression: the coordinate descent method (glmnet) developed by Simon et al. (2011), the predictor-corrector (glmpath) introduced by Park et al. (2007) and the gradient ascent algorithm (penalized) proposed by Goeman (2010). In our simulation study we generate survival times t_i, $i = 1, 2, \ldots, n$, following exponential distributions with subject-specific parameters $\lambda_i = exp(\beta^T X_i)$. In the linear form $\beta^T X_i$, the variables $X_1, \ldots, X_p$ are sampled from a multivariate normal density $N(\mathbf{0}, \Sigma)$ where the entries of Σ are fixed to $corr(X_j, X_k) = \rho^{|j-k|}$ for $\rho \in (0.5, 0.7, 0.9)$. The censorship is randomly assigned to the survival times with probability $\pi \in (0.2, 0.4, 0.6)$. We fix the sample size n to 50 and the number of predictors p to 100 to emulate an scenario in which $p > n$. From the 100 predictors used, we fix first 30 to 2 and the remaining 70 are set to zero.

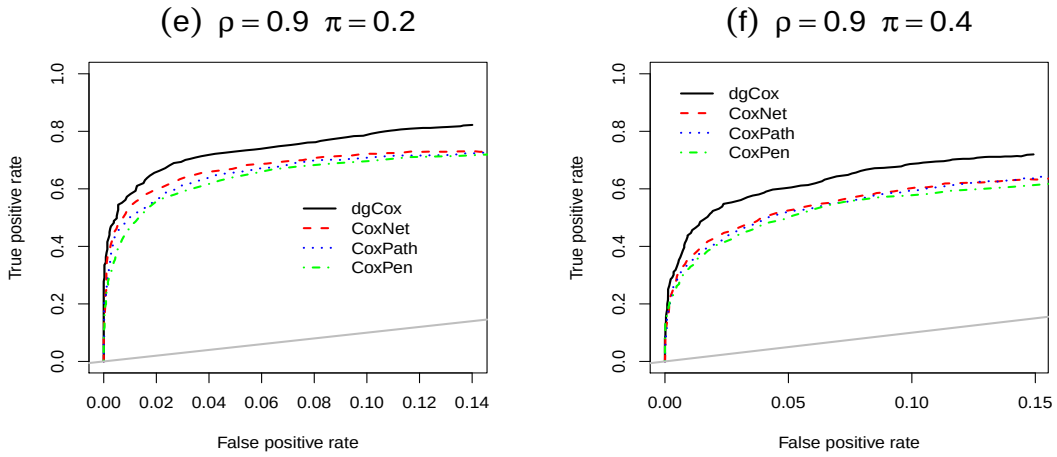

Figure 1: Results from the simulation study; for each scenario we show the averaged ROC curve (using 100 datasets) for the dgCox, the coordinate descent method (glmnet), the predictor-corrector (glmpath) and the gradient ascent algorithm (penalized). The 45-degree diagonal is also included in the plots.

For each one of the previous scenarios we generate 100 datasets and we calculate the receiver operating characteristic (ROC) curves for the four methods. In scenarios (a) and (b), where $\rho = 0.5$, the four method methods exhibit a similar performance, having overlapping curves for both levels of censorship. A similar performance of the methods has been also observed for combinations of smaller values of ρ and π. In scenarios (c) and (d), where the value of ρ increases to 0.7, the glmnet, glmpath and penalized approaches still overlap, whereas the dgCox model appears to be consistently the best method. In scenarios (e) and (f) where the correlation among the predictors is high, say $\rho = 0.9$, the 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 the glmnet, glmpath and penalized approach. The results for scenario (e) and (f) are shown as a ROC curve in Figure 1.

In summary, the dgCox method shows a better or equal behaviour than the glmnet, glmpath and penalized approaches in all the scenarios. In particular, we observe that the performance of the four methods is similarly affected by the inclusion of different proportions of censored data. However, the dgCox models is much more efficient in cases in which the predictor variables show high correlations levels.

4 Case study: large-B-cell lymphoma

In this section we analyze the diffuse large-B-cell lymphoma dataset (DLBCL) originally presented by Rosenwald and et. al. (2002). This dataset consist on the gene expression measurements and survival times of 240 patients in 7399 genes after a chemotherapy treatment. Originally, the purpose of the experiment was to formulate a molecular predictor model of survival after the chemotherapy treatment. Our goal is to use the dgLars model to identify the most influential genes in such predictor.

From the original 240 patients, we first remove the five with survival time zero. We impute the missing data using the k-nearest neighbour approach proposed in Troyanskaya et al. (2001). In order to test the robustness of the estimations, we use the original partition of the data into a training set, consisting on 156 patients and a validation set consisting of 79 patients.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Augugliaro, L., Mineo, A., and Wit, E. (2013). Differential geometric least angle regres-

sion: a differential geometric approach to sparse generalized linear models. *Journal of the Royal Statistical Society. Series B*, 75(3):471–498.

Burbea, J. and Rao, R. C. (1982). Entropy Differential Metric, Distance and Divergence Measures in Probability Spaces - A Unified Approach. *Journal of Multivariate Analysis*, 12:575–596.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34(2):187–220.

Cox, D. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

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

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99.

Goeman, J. J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, 52(1):70–84.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, London, 2nd edition.

Moolgavkar, S. 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(1):346–359.

Park, Young, M., Hastie, and Trevor (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.

Prentice, R. and Self, S. (1983). Asymptotic distribution theory for cox-type regression models with general relative risk form. *The Annals of Statistics*, 11(3):804–813.

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

Rosenwald, A. and et. al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*, 346(25):1937–1947.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Thomas, D. (1977). Addendum to the paper by Liddell, McDonald, Thomas and Cunliffe. *Journal of the Royal Statistical Society. Series A*, 140(4):483–485.

Thomas, D. (1981). General relative-risk models for survival time and matched case-control analysis. *Biometrics*, 37(4):673–686.

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

Troyanskaya, O. G., Cantor, M., Sherlock, G., Brown, P. O., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.