Regularization in Cox Frailty Models

Andreas Groll[1], Trevor Hastie[2], Gerhard Tutz[3]

[1] Ludwig-Maximilians-Universität Munich, Department of Mathematics, Theresienstraße 39, 80333 Munich, Germany
[2] University of Stanford, Department of Statistics, 390 Serra Mall, Sequoia Hall, California 94305, United States
[3] Ludwig-Maximilians-Universität Munich, Department of Statistics, Akademiestraße 1, 80799 Munich, Germany

E-mail for correspondence: groll@math.lmu.de

Abstract: In all sorts of regression problems it has become more and more relevant to face high dimensional data with lots of potentially influential covariates. A possible solution is to apply estimation methods that allow to select the relevant covariates. These methods are often based on suitable penalization of the corresponding regression models likelihood function. In this work, a penalization approach for variable selection in a particular regression model for survival analysis is considered, the so-called Cox frailty model. As in many applications the influence of some covariates changes over time, also time-varying effects are considered. A suitable penalization approach then has to cover several model selection issues. Besides, the method incorporates a multiplicative log-normal frailty distribution, resulting in flexible and sparse hazards models for modeling survival data.

Keywords: Variable selection; LASSO; Cox frailty model; Time-varying coefficients; Penalization.

1 Introduction

Among the class of models designed for continuous event times, the proportional hazards models play a major role, in particular the famous Cox model (Cox, 1972). The Cox model assumes the semi-parametric hazard

$$\lambda_i(t|\boldsymbol{x_i}) = \lambda_0(t)\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}),$$

where $\lambda_i(t|\boldsymbol{x_i})$ is the hazard for observation i at time t, conditionally on the covariates $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. $\lambda_0(t)$ is the shared baseline hazard, and $\boldsymbol{\beta}$ the fixed effects vector. Note that in the continuous time case the hazard rate $\lambda_i(t|\boldsymbol{x})$ is defined as $\lambda_i(t|\boldsymbol{x}) = \lim_{\Delta t \to 0} P(t \leq T < \Delta t | T \geq t, \boldsymbol{x})/\Delta t$, representing the instantaneous risk of a transition at time t. Inference is usually based on maximization of the corresponding partial likelihood. This

approach allows estimation of $\boldsymbol{\beta}$ while ignoring $\lambda_0(t)$ and performs well in classical problems with more observations than predictors. To combat the $p > n$ problem, Tibshirani (1997) proposed the use of the so-called least absolute shrinkage and selection operator (LASSO) penalty in the Cox model. Since then, several extensions have been proposed, compare Park and Hastie (2007) or Goeman (2010), just to mention two.

2 Cox frailty model with time-varying coefficients

If dependencies within clusters of observations exist or if there is heterogeneity between clusters, these can be captured effectively by frailty models. However, parameter estimation in frailty models is more challenging than in the Cox model, since the corresponding profile likelihood does not have a closed form solution. In the Cox proportional hazards frailty model the hazard rate of the j-th subject belonging to subgroup-cluster i, conditionally on the covariates $\boldsymbol{x}_{ij}$ and the shared frailty u_i, is given by

$$\lambda_{ij}(t|\boldsymbol{x}_{ij}, b_i) = b_i \lambda_0(t) \exp(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}),$$

where the frailties $b_i, i = 1, ...n$, are frequently assumed to follow a gamma distribution because of its mathematical convenience.
While there exist several R packages to fit Cox frailty models, for example `frailtypack` (Rondeau et al., 2012) and `survival` (Therneau, 2014), only limited approaches to variable selection are yet available. Though Fan and Li (2002) as well as Androulakis et al. (2012) have contributed considerable works in this context, no software implementation is yet available.
While some multiplicative frailty distributions, such as e.g. the gamma and the inverse Gaussian, have already been extensively studied (compare Androulakis et al., 2012) and closed form representations of the log-likelihoods are available, in some situations the log-normal distribution is more intuitive and besides, it generally allows for more flexible and complex predictor structures though the corresponding model is computationally more demanding. The conditional hazard function with multiplicative frailties following a multivariate log-normal distribution yields in its general form

$$\lambda_{ij}(t|\boldsymbol{x}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \lambda_0(t) \exp(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{u}_{ij}^{\mathrm{T}} \boldsymbol{b}_i),$$

where the random effects follow a multivariate Gaussian distribution, i.e. $\boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{Q}(\boldsymbol{\theta}))$, with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{Q}(\boldsymbol{\theta})$, which is depending on a vector of unknown parameters $\boldsymbol{\theta}$. In this case, a penalized quasi-likelihood (PQL) approach based on Laplace approximation can be used for estimation, following Breslow and Clayton (1993) in their approach for the generalized linear mixed model (GLMM). In this context, it is especially important to provide effective estimation algorithms, as standard procedures for determination of tuning parameters such as cross validation are usually very time-consuming.

While for Cox frailty models with the simple predictor structure $\eta_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{u}_{ij}^{\mathrm{T}}\boldsymbol{b}_i$ in the hazard function some solutions have already been given (see e.g. Fan and Li, 2002, and Androulakis et al., 2012), often more complex structures of the linear predictor need to be taken into account. For example, time-varying effects $\gamma_k(t)$ can be incorporated into the linear predictor. For observation i from cluster j, this yields the hazard rate

$$\lambda_{ij}(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \lambda_0(t) \exp\left(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k=1}^{s} z_{ijk}\gamma_k(t) + \boldsymbol{u}_{ij}^{\mathrm{T}}\boldsymbol{b}_i\right).$$

A standard way to estimate the time-varying effects $\gamma_k(t)$ is to expand them in equally spaced B-splines yielding $\gamma_k(t) = \sum_{m=1}^{m_k} \alpha_{k,m} B_{k,m}(t; d)$, where $\alpha_{k,m}, m = 1, \ldots, m_k$, denote unknown spline coefficients, which need to be estimated, and $B_{k,m}(t; d)$ is the m-th B-spline basis function of the k-th time-varying effect of degree d. For a detailed description of B-splines, see for example Wood (2006) and Ruppert et al. (2003).
In general, for the cumulative baseline hazard $\Lambda_0(\cdot)$ often the "least informative" nonparametric modeling is considered. More precisely, with $t_1^0 < \ldots. < t_N^0$ denoting the observed event times, the least informative nonparametric cumulative baseline hazard $\Lambda_0(t)$ has a possible jump h_j at every observed event time t_j^0, i.e. $\Lambda_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$. However, the estimation procedure may be stabilized, if instead, similar to the time-varying effects, a semi-parametric baseline hazard is considered, which can be flexibly estimated within the B-spline concept. Then, using the transformation $\gamma_0(t) := \log(\lambda_0(t))$ and setting $z_{ij0} = 1$ for all i, j, we can specify the hazard rate as

$$\lambda_{ij}(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \exp\left(\eta_{ij}(t)\right), \tag{1}$$

with $\eta_{ij}(t) := \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k=0}^{s} z_{ijk}\left(\sum_{m=1}^{m_k} \alpha_{k,m} B_{k,m}(t; d)\right) + \boldsymbol{u}_{ij}^{\mathrm{T}}\boldsymbol{b}_i$. In general, the estimation of parameters in the predictor (1) can be based on Cox's well-known full log-likelihood, which is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}) = \sum_{i=1}^{n}\sum_{j=1}^{N_i} \delta_{ij}\eta_{ij}(t_{ij}) - \int_0^{t_{ij}} \exp(\eta_{ij}(s))ds, \tag{2}$$

where n denotes the number of clusters, N_i the cluster sizes and the survival times t_{ij} being complete if $\delta_{ij} = 1$ and right censored if $\delta_{ij} = 0$.

3 Penalization

Note that certain questions of model selection are related to the type of predictor (1). In particular, one has to determine which covariates should be included in the model, or, which of the covariates included have a time-varying effect. So our objective is to develop a penalization approach for

variable selection in Cox frailty models with time-varying coefficients, such that single varying effects are either included, are included in the form of a constant effect or are totally excluded. These model selection issues can be achieved by incorporating a suitable penalty into the fitting procedure. We propose to subtract the following penalty from the Cox frailty log-likelihood

$$\xi \cdot J_\zeta(\boldsymbol{\alpha}) = \xi \left(\zeta \sum_{k=1}^{s} \psi_k ||(\vartheta_{k,2}, \ldots, \vartheta_{k,m_k})||_2 + (1-\zeta) \sum_{k=1}^{s} \phi_k ||(\alpha_{k,1}, \ldots, \alpha_{k,m_k})||_2 \right),$$

where $||\cdot||_2$ denotes the L_2-norm, $\xi \geq 0$ and $\zeta \in (0,1)$ are tuning parameters and $\vartheta_{k,l} = \alpha_{k,l} - \alpha_{k,l-1}$. The first term of the penalty controls the smoothness of the time-varying covariate effects, whereby for values of ξ and ζ large enough, all differences $\alpha_{k,l} - \alpha_{k,l-1}, l = 2, ..., m_k$, are removed from the model, resulting in constant covariate effects. As the B-splines of each varying coefficient sum up to one, a constant effect is obtained, if all spline coefficients are equal. Hence, the first penalty term does not affect the spline's global level. The second term penalizes all spline coefficients belonging to a single time-varying effect in the way of a group LASSO and, hence, controls selection of covariates. Both tuning parameters ξ, ζ should be chosen by an appropriate technique, such as for example by K-fold cross validation. The terms $\psi_k := \sqrt{m_k - 1}$ and $\phi_k := \sqrt{m_k}$ represent weights assigning different amounts of penalization to different parameter groups, relative to the respective group size. Within the estimation procedure, i.e. the corresponding Newton-Raphson algorithm, local quadratic approximations of the penalty term are used, following Oelker and Tutz (2013). Note that the penalty from above may be easily extended by a conventional LASSO penalty for time-constant fixed effects $\beta_k, k = 1, ..., p$.

Besides, as also the baseline hazard in the predictor (1) is considered to be semi-parametric, the penalty from above should be further extended by adding another penalty term to control the roughness of the baseline. If the smooth log-baseline hazard $\gamma_0 = \log(\lambda_0(t))$ is twice differentiable, one could for example penalize its second order derivatives, similar to Yu et al. (2012). Alternatively, if $\gamma_0(t)$ is once again expanded in B-spline basis functions, i.e. $\gamma_0(t) = \sum_{m=1}^{m_0} \alpha_{0,m} B_{0,m}(t; d)$, simply the squared differences of adjacent spline weights $\alpha_{0,l}$ and $\alpha_{0,l-1}, l = 2, \ldots, m_0$, could be penalized. Hence, beside $\xi \cdot J_\zeta(\boldsymbol{\alpha})$, also the penalty term

$$\xi_0 \cdot J_0(\boldsymbol{\alpha_0}) = \xi_0 \left(\sum_{l=2}^{m_0} (\alpha_{0,l} - \alpha_{0,l-1})^2 \right)$$

has to be subtracted from the Cox frailty log-likelihood, with the vector $\boldsymbol{\alpha_0}^\mathrm{T} := (\alpha_{0,1}, \ldots, \alpha_{0,m_0})$ collecting those spline coefficients from $\boldsymbol{\alpha}$ that correspond to the baseline hazard. Although this adds another tuning parameter ξ_0, it turns out that in general it is not worthwhile to select also ξ_0 on a grid of possible values. Note here that we have already obtained similar findings with regard to penalization of the baseline hazard in discrete

frailty survival models, see Tutz and Groll (2014). While probably some care should be taken to select ξ and ζ, which determine the performance of the selection procedure, the estimation procedure is already stabilized in comparison to the usage of the least informative nonparametric cumulative baseline hazard $\Lambda_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$ for a moderate choice of ξ_0.
As already mentioned in Section 2, a possible strategy to maximize the full log-likelihood (2) is based on the PQL approach, which was originally suggested for GLMMs by Breslow and Clayton (1993). Typically, the covariance matrix $\boldsymbol{Q}(\boldsymbol{\theta})$ of the random effects $\boldsymbol{b}_i$ depends on an unknown parameter vector $\boldsymbol{\theta}$. Hence, the joint likelihood-function can be specified by the parameter vector of the covariance structure $\boldsymbol{\theta}$ and parameter vector $\boldsymbol{\delta}^{\mathrm{T}} := (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{b}^{\mathrm{T}})$. The corresponding *marginal* log-likelihood then yields

$$l^{mar}(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left(\int L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}_i) p(\boldsymbol{b}_i | \boldsymbol{\theta}) d\boldsymbol{b}_i \right),$$

where $p(\boldsymbol{b}_i|\boldsymbol{\theta})$ denotes the density function of the random effects and the quantities $L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}_i) := \prod_{j=1}^{N_i} \exp(\eta_{ij}(t_{ij}))^{\delta_{ij}} \exp\left(-\int_0^{t_{ij}} \exp(\eta_{ij}(s))ds\right)$ represent the likelihood contributions of single clusters $i, i = 1, \ldots, n$. Approximation along the lines of Breslow and Clayton (1993) yields

$$\begin{aligned} l^{app}(\boldsymbol{\delta}, \boldsymbol{\theta}) &= \sum_{i=1}^{n} \log L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{b}_i) - \frac{1}{2} \boldsymbol{b}^{\mathrm{T}} \boldsymbol{Q}(\boldsymbol{\theta}) \boldsymbol{b} \qquad (3) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{N_i} \left(\delta_{ij} \eta_{ij}(t_{ij}) - \int_0^{t_{ij}} \exp(\eta_{ij}(s))ds \right) - \frac{1}{2} \boldsymbol{b}^{\mathrm{T}} \boldsymbol{Q}(\boldsymbol{\theta}) \boldsymbol{b}, \end{aligned}$$

the penalty term $\boldsymbol{b}^{\mathrm{T}} \boldsymbol{Q}(\boldsymbol{\theta}) \boldsymbol{b}$ resulting from the approximation based on the Laplace method. The PQL approach usually works within the profile likelihood concept. It is distinguished between estimation of $\boldsymbol{\delta}$, given the plug-in estimate $\hat{\boldsymbol{\theta}}$ and resulting in profile likelihood $l^{app}(\boldsymbol{\delta}, \hat{\boldsymbol{\theta}})$, and estimation of $\boldsymbol{\theta}$.

4 Estimation

Estimation is now based on maximization of the penalized log-likelihood, which is obtained by expanding the approximate log-likelihood $l^{app}(\boldsymbol{\delta}, \boldsymbol{\theta})$ from (3) to include the penalty terms $\xi_0 \cdot J_0(\boldsymbol{\alpha_0})$ and $\xi \cdot J_\zeta(\boldsymbol{\alpha})$, i.e.

$$l^{pen}(\boldsymbol{\delta}, \boldsymbol{\theta}) = l^{app}(\boldsymbol{\delta}, \boldsymbol{\theta}) - \xi_0 \cdot J_0(\boldsymbol{\alpha_0}) - \xi \cdot J_\zeta(\boldsymbol{\alpha}).$$

The estimation procedure is based on a conventional Newton-Raphson algorithm, while local quadratic approximations of the penalty term are used, following Oelker and Tutz (2013).

It turns out that the combination of the proposed penalization approach for variable selection in Cox frailty models with time-varying coefficients with the promising class of multivariate log-normal frailties results in very flexible and sparse hazard rate models for modeling survival data.

References

Androulakis, E., Koukouvinos, C., and Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine*, **31**, 2223 – 2239.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association*, **88**, 9 – 25.

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

Fan, J. and Li, R. (2002). Variable selection for Coxs 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**, 70 – 84.

Oelker, M.-R. and Tutz, G. (2013). A General Family of Penalties for Combining Different Types of Penalties in Generalized Structured Models. *Technical Report*, **139**, Department of Statistics, LMU Munich.

Park, M.Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society*, **Series B 19**, 659 – 677.

Ruppert, D., Wand, M.P. , and Carroll,R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, **16**, 385 – 395.

Tutz, G. and Groll, A. (2014). Variable Selection in Discrete Survival Models Including Heterogeneity. *Technical Report* **167**, Department of Statistics, LMU Munich.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall.

Yu, Z., Lin, X., and Tu, W. (2012). Semiparametric Frailty Models for Clustered Failure Time Data. *Biometrics* **68(29)**, 429–436.