Universidade de Vigo

Model Building in Non Proportional Hazard Regression

Mar Rodríguez-Girondo, Thomas Kneib,
Carmen Cadarso-Suárez and Emad Abu-Assi

Report 12/09

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://webs.uvigo.es/depc05/
E-mail: depc05@uvigo.es

Universidade de Vigo

Model Building in Non Proportional Hazard Regression

Mar Rodríguez-Girondo, Thomas Kneib,
Carmen Cadarso-Suárez and Emad Abu-Assi

Report 12/09

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita:

Universidade de Vigo
Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

Model Building in Non Proportional Hazard Regression

Mar Rodríguez-Girondo[1,2], Thomas Kneib[3], Carmen Cadarso-Suárez[2] and Emad Abu-Assi[4]

[1]SiDOR Research Group, University of Vigo, Spain

[2]Department of Statistics, University of Santiago de Compostela, Spain

[3]Center for Statistics, Georg August University, Göttingen, Germany

[4]Cardiology Department, Hospital Clínico de Santiago de Compostela, Spain

Abstract

Recent developments of statistical methods allow for a very flexible modeling of covariates affecting survival times via the hazard rate, including also the inspection of possible time-dependent associations. Despite their immediate appeal in terms of flexibility, these models typically introduce additional difficulties when a subset of covariates and the corresponding modeling alternatives have to be chosen, i.e. for building the most suitable model for given data. This is particularly true when potentially time-varying associations are given. We propose to conduct a piecewise exponential representation of the original survival data to link hazard regression with estimation schemes based on the Poisson likelihood to make recent advances for model building in exponential family regression accessible also in the non proportional hazard regression context. A two-stage stepwise selection approach, an approach based on doubly penalized likelihood and a componentwise functional gradient descent approach will be adapted to the piecewise exponential regression problem and compared via an intensive simulation study. An application to prognosis after discharge for patients who suffered a myocardial infarction supplements the simulation to demonstrate the pros and cons of the approaches in real data analyses. A software implementation is also provided.

Keywords: *variable selection; survival analysis; boosting; generalized additive models; piecewise exponential model; penalized likelihood; model choice.*

1 Introduction

Development of model building strategies in survival analysis is an important methodological issue relevant in many biomedical applications. Prognostic models are increasingly used to complement physician judgment for decision making in clinical practice. A critical step when developing multivariate risk prediction models is to decide which are the predictors that should be included and which ones could be neglected in order to get an accurate but interpretable model. The increasing

availability of potentially relevant information on patients and the complex relations among them makes it difficult for practitioners to elucidate which are the best combinations of predictors in order to reasonable risk stratification of patients. Moreover, continuous covariates may require nonlinear modeling if a purely linear predictor is not sufficient to capture complex association structures or covariates may impact survival in a time-varying manner. This multitude of modeling alternatives (in addition to other problems such as censoring and non-standard likelihoods in hazard regression) makes model building in the survival context a challenging yet very relevant task.

In this paper, we follow the classical Cox proportional hazards framework in modeling the hazard rate of the survival time in terms of covariates but extend the classical framework in two important directions: the effect of covariates may vary over time and more complex, nonlinear relationships between covariates and survival are allowed. More specifically, to analyze the impact of a set of covariates collected in the vector $\mathbf{x}$, consider the following smooth general specification of the hazard function:

$$\lambda(t|\mathbf{x}) = \exp(\eta(t|\mathbf{x})), \tag{1}$$

with additive predictor of the form

$$\eta(t|\mathbf{x}) = g_0(t) + \sum_{u=1}^{U} x_u\beta_u + \sum_{v=1}^{V} s_v(x_v) + \sum_{w=1}^{U+V} x_w g_w(t)) \tag{2}$$

where β_u refers to the traditional linear effects, s_v are smooth effects of continuous covariates and g_w are smooth time-varying effects of both categorical and continuous covariates. The log-baseline g_0 is included in the predictor and will also be modeled by a smooth function included in the estimation. Although a lot of attention was devoted in the last years to accommodate both smooth and time-varying effects in survival regression (see [3] for a recent review), model building strategies including variable selection in such complex survival model settings are still a big methodological challenge.

In summary, when developing a prognostic model, we are not only concerned about the correct functional form of the studied covariates, but instead we also aim to build an interpretable model, where only the relevant prognostic factors are included with a correct level of complexity, i.e., we aim at a compromise between goodness of fit and parsimony. Specifically, we are interested in answering questions such as:

- Should a covariate enter the specified hazard model or is it not important and can be ignored?
- Should a continuous covariate enter the model as a purely linear effect or is more smoothness indeed required?

- Is the effect of a covariate constant along the follow-up time or does its effect vary over time?

Several recent contributions to variable selection and model choice within the exponential family regression framework of generalized additive models (GAMs) are available (see [17] and references therein for an overview) but these are typically not immediately applicable in hazard regression. Broadly speaking, we may primarily distinguish between two different approaches to the model building problem in GAM settings: (i) Multi-stage procedures conducting a sequential comparison of different model alternatives based on the comparison of some information criteria or hypothesis testing procedures, and (ii) one-step algorithms. The latter deal with model building by means of regularization of effects, combining in a single step model fitting, variable selection and model choice. Within this general regularization framework, we find very different candidates such as direct shrinkage methods based on penalized maximum likelihood estimation [17] but also procedures achieving regularization indirectly such as functional gradient descent boosting [24, 4, 15] originating from machine learning.

As already noted before, specific contributions in model building for flexible survival models [2, 23] are scarcer than in the GAM context, especially when referring to one-step methods. Handling Cox-type maximum likelihood, in particular when time-varying effects are present, requires specialized estimation schemes that are often highly computationally demanding and less stable than in standard regression settings. As a consequence, the application of existing model building methods developed in standard regression settings is not straightforward in the survival context. However, the gap between GAMs and additive survival models can be closed by means of a piecewise exponential representation of the original data relying on a piecewise constant approximation to all time-dependent quantities. In practice, the piecewise exponential representation implies some data augmentation but allows for convenient estimation based on a Poisson-likelihood which also enables to adapt model building methods proposed for GAMs to hazard regression.

In summary, the aim of this paper is twofold: On the one hand, we will present the piecewise exponential representation of survival models as a flexible alternative to Cox-type estimation and, on the other hand, we will show the convenience of this approach when model selection is the ultimate goal of the analysis. In addition, we aim to compare three different approaches for model building originating from different conceptual fields

- a stepwise multi-stage approach based on Akaike's information criterion [11];
- a shrinkage method for generalized additive models based on penalized maximum likelihood estimation [17];

- and a Poisson-likelihood boosting algorithm [4, 15].

In all cases, we use a representation of the smooth effects in terms of penalized splines [7].

As application, we aim to build a flexible and parsimonious model for predicting survival after myocardial infarction, conducting data-driven variable selection among several potentially influential predictors.

The rest of this paper is organized as follows: In section 2, flexible additive survival models are revisited and a piecewise exponential representation is proposed. Penalized spline smoothing is considered for representing both nonlinear effects of continuous covariates and time-varying effects. Section 3 presents three different model building alternatives in this context which are compared by means of a simulation study in Section 4. Section 5 presents an application to cardiology data and, finally, the paper is concluded with a discussion section.

2 Flexible Models for Survival Data

2.1 Structured Hazard Regression Models

Let the observations be given by $(t_i, \delta_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, where t_i is the right-censored survival time, δ_i the non-censoring indicator and $\mathbf{x}_i = (x_{1i}, \ldots, x_{qi})'$ a vector of q covariates which may affect survival. Denoting by $\mathbf{z}_i = (t_i, x_{1i}, \ldots, x_{ni})'$ the vector containing both observed time and covariates, the additive hazard regression model introduced in (1) and (2) can be synthetically rewritten as [10, 14]:

$$\lambda(t_i|\mathbf{z}_i) = \exp(\eta(\mathbf{z}_i)), \quad i = 1, \ldots, n \tag{3}$$

with structured additive predictor of the form

$$\eta_i(\mathbf{z}_i) = \beta_0 + \sum_{j=1}^{J} f_j(\mathbf{z}_i). \tag{4}$$

The functions f_j are a generic representations of different types of covariate effects that will enable for a compact and convenient presentation of estimation and model building approaches in the rest of this paper. Specifically, we will consider the following possible effects for functions f_j: (i) Linear effects $f_j(x) = f_{linear}(x) = x\beta$, where x is a categorical covariate of the vector $\mathbf{z}$. (ii) Nonparametric, smooth effects $f_j(x) = f_{smooth}(x)$, where x is a continuous covariate. (iii) Time-varying effects $f_j(x) = xf_{smooth}(t)$, where x can be either continuous or categorical. The baseline hazard rate is also included in the linear predictor by reparametrizating it as $f_j(t) = f_{smooth}(t) = \log(\lambda_0(t))$. Note that the inclusion of baseline hazard as a smooth effect as any other continuous covariate besides of

being of interest from a practical point of view, allows to use the full likelihood for model estimation and prediction.

In this framework and under the usual assumptions about non-informative censoring and conditional independence between the observations, the full data log-likelihood can be expressed as

$$\ell = \sum_{i=1}^{n} [\delta_i \eta_i(\mathbf{z}_i) - \int_0^{t_i} \lambda_i(t) dt] \tag{5}$$

Note that expression (5) involves an integral over the hazard rate (potentially including time-varying effects) and hence numerical integration would be required to actually evaluate the log-likelihood and its derivatives. This is not easily handled when models become complicated and it may frequently lead to numerical problems as well as instable estimates in hazard regression situations. Hence, the direct approach to estimating hazard regression models is not the preferable one when aiming at model selection that may require the estimation of complex models and/or a large number of different models.

2.2 Piecewise Exponential Representation

We therefore now introduce a modification of the initial data set by means of data augmentation which enables a representation of the survival model given in (3) in terms of a piecewise exponential model [8]. The main advantage of such a modification is that it allows for an equivalent and more convenient Poisson likelihood based estimation scheme that links hazard regression with exponential family regression and the corresponding model building concepts. The key result relies on the fact that treating deaths as Poisson conditional on exposure times leads to exactly the same likelihood and therefore also the same estimates as treating the exposure times as censored observations from an exponential distribution with piecewise constant hazard rate.

The basic idea of the piecewise exponential model is therefore that the proportional hazard assumption holds at least over short periods of time such that all time-varying effects can be treated as piecewise constant. Denoting by a_r the largest of all observed times, the time axis can be divided into a sequence of short intervals $[a_0, a_1)$, $[a_1, a_2)$,..., $[a_{r-1}, a_r)$, $[a_r, \infty)$ where the hazard is assumed to be constant for times t within the intervals $[a_{m-1}, a_m)$, $m = 1, \ldots, r$. This means that the data set has to be augmented in such a way that for every individual i, $i = 1, \ldots, n$, we obtain an observation row for each interval $[a_{m-1}, a_m)$ beginning with the first one up to the interval in that observation time t_i falls. The modified data set then also contains some new variables: instead of the non-censoring indicator δ_i, the new data set contains y_{im}, an indicator of the interval where the event occurs, and instead of the observed survival time t_i, t_{im}, the time lived by the individual i in the m-th interval and the variable a considered as a continuous variable in the new augmented

data set. The rest of the covariates values are replicated in each interval. To fix ideas, we specify the data augmentation process with a small example. Suppose that we consider an equidistant grid with interval width 1 and the first two observations of the original data set are given by:

id	t	δ	x_1
1	3.5	1	2
2	1.2	0	1

The corresponding augmented data set is as follows:

id	y	a	Δ	x_1
1	0	1	log(1)	2
1	0	2	log(1)	2
1	0	3	log(1)	2
1	1	4	log(0.5)	2
2	0	1	log(1)	1
2	0	2	log(0.2)	1

Note that the data augmentation scheme also routinely allows to include time-varying covariates by assigning different time-dependent covariate values to the intervals in the augmented data set.

Despite being parametric in a strict sense, the piecewise exponential model can still be considered a nonparametric model due to its large flexibility when a large number of intervals is employed. The log-likelihood expression (5) for the original hazard regression model turns to

$$\ell_{pem} = \sum_{m=1}^{r} \sum_{i \in R_m} (y_{im}\eta_{im} - \Delta_{im} + \exp(\eta_{im})) \tag{6}$$

for the piecewise exponential model, where R_m is the risk set for $[a_{m-1}, a_m)$, i.e. R_m contains the indices of those individuals that have not experienced an event before a_{m-1}, and $\Delta_{im} = \log(t_{im})$ is an offset term. Note that since log-baseline hazard and time-varying effects turned constant in each interval, i.e., $f_{smooth}(t) = f_{jm}$ is a constant quantity over $[a_{m-1}, a_m)$, the time dependence has no influence in η_{im}.

It now turns out that the log-likelihood given in (6) is equivalent to a Poisson log-likelihood for responses y_{im} with offset Δ_{im} [12, 16] and thus, the estimation can be performed using standard generalized additive regression with a Poisson error structure. Specifically, the new set of observations consists of $(y_i, \mathbf{z}_i)$, $i = 1, \ldots, n'$ where y is the response variable, $\mathbf{z}_i = (a_i, x_{1i}, \ldots, x_{qi}, \Delta_i)$ as vector of covariates and n' is the number of observations resulting from the data augmentation explained above. Hence, the conditional expectation of y_i is related to the covariates via

$$E(y_i|\mathbf{z}_i) = \exp(\eta(\mathbf{z}_i) + \Delta_i) \tag{7}$$

The specification of the predictor is therefore equivalent to the one given in (4), while Δ acts as an offset term that is assigned a fixed coefficient of 1 to adjust observations with respect to the survival

time experienced in a given interval. Within this representation, the original time-varying effects in (3) are considered as standard varying coefficient terms $f_j(x) = x f_{smooth}(a)$ [9]. From now on, inferential results and variable selection strategies will be based on the resulting Poisson model given by (7).

2.3 Penalized Spline Smoothing

In the following, all smooth functions f_{smooth} of covariates $a, x_i, \ldots, x_q$ included in (7) will be modeled using penalized splines [7]. Thus, the nonparametric problem is replaced by a parametric equivalent, in which a vector of regression coefficients is estimated under a smoothness penalty. The general idea is to approximate the functions f_j by linear combinations of B-spline basis functions, i.e.

$$f_j(x) = \sum_{j=1}^{d_k} \beta_j B_j(x) \tag{8}$$

where β_j are the regression coefficients corresponding to the B-spline basis defined over a grid of d_k equally spaced knots. As a result, we can express each of the predictor components as the product of an appropriate design matrix $\mathbf{Z}_j$ and a vector β_j. Moreover, a penalty term is added to control the level of smoothness by penalizing wiggly functions when estimating β_j. The most commonly used penalization term is based on the integral of the second derivative of the smooth functions, f_j, i.e.

$$\text{pen}(f_j) = \frac{1}{2}\lambda_j \int_0^\infty [f_j''(x)]^2 dx. \tag{9}$$

Since equation (9) is a quadratic form in the vector of regression coefficients β_j, it can be written as $\frac{1}{2}\lambda_j \beta_j' \mathbf{K}_j \beta_j$, where the penalty matrix $\mathbf{K}_j$ is a positive semidefinite matrix and $\lambda_j \geq 0$ a smoothing parameter. Since the smooth functions for the nonlinear effects are represented in terms of B-splines it allows to approximate the penalty term in terms of the squared differences of coefficients associated with adjacent basis functions [7]. As a result, the difference penalty matrix can be written as $\mathbf{K}_j = \mathbf{D}'\mathbf{D}$, where $\mathbf{D}$ is the second order difference matrix of neighboring coefficients.

3 Model Building Strategies

In this section, we propose three methods for model building in the context of additive hazard regression with smooth time-dependent effects utilizing the equivalent piecewise exponential representation in terms of a Poisson generalized additive model. More specifically, we consider an adaptation of the multi-stage stepwise approach based on Akaikes information criterion proposed for flexible Cox-type models by [11], the double penalty approach for model building in generalized

additive models developed by [17], and a functional gradient descent boosting approach for Poisson regression following [4, 15].

Although alternative approaches (as the ones discussed in [17]) could also be adapted to hazard regression, we focus on these three approaches for the following reasons:

- The multi-stage stepwise procedure is a representative for classical stepwise model selection algorithms specifically adapted to additive hazard regression with non proportional hazards. Its performance therefore allows to investigate whether stepwise approaches are competitive to more advanced and automatic approaches such as shrinkage methods.
- The double penalty approach is a natural extension of penalized likelihood estimation for GAMS that allows to completely drop specific terms from the model. It is therefore conceptually close to classical GAM estimation and may therefore be especially attractive for practitioners already familiar with this framework.
- Boosting was chosen as a representative of modern regression approaches originating from the machine learning community because it also allows to estimate structured models that can be interpreted in complete analogy to classical GAM models. Since our goal is to investigate model building and not to optimize predictive performance, boosting is a natural candidate as compared to for example survival random forests.
- All three approaches rely on the same principles (penalized spline specifications of the nonlinear and time-varying effects, additive hazard regression specification, Poisson likelihood arising from the PEM representation) and can therefore be directly compared when using, for example, the same specifications for the spline bases. We therefore do no have to worry about suitably choosing specific tuning constants to enable a fair comparison between the approaches.

We will discuss some possible alternative routes for model building in survival models in the final section of this paper.

3.1 Two-Stage Stepwise Selection

The two stage stepwise procedure considered in the following is an adaptation of the original proposal for building Cox-type structured hazard models [11] to exponential family additive regression to fit with the PEM representation of the additive hazard regression model. The method combines variable and model selection in a multi-step fashion relying on a chosen optimality criteria. We will consider the conditional AIC (cAIC) in the following but other criteria can be chosen according to the goal of the analysis.

The basic structure of the two-stage stepwise selection algorithm is as follows:

(i) Given a working model (which in the beginning usually will be the empty model without any covariates) define a set of modeling alternatives for each covariate not already included in the model. In our case, these alternatives will be linear effect versus time-varying effect for categorical variables and linear effect versus nonparametric effect versus time-varying effect for continuous.

(ii) All models arising from adding a new covariate in a specific modeling alternative to the working model are estimated and the corresponding cAIC is recorded.

(iii) If none of the fitted models in step (ii) improves the current model, the algorithm is terminated. Otherwise the model with the lowest cAIC is selected as the new working model and the corresponding covariate is deleted from the choice set of candidate variables.

(iv) Finally, a backward deletion on the current model is performed by dropping one covariate at a time. If an improvement on cAIC is achieved, the reduced model is then the working model and the corresponding variable will again be included in the choice set of candidate variable to enter the model.

(v) Go back to (i).

In each step, models are fitted by maximizing penalized log-likelihood derived from the representation of the smooth effects f_j in terms of P-splines, i.e. by maximizing

$$\boldsymbol{l_{\mathbf{pen}}(\beta) = l(\beta) - \sum_{j=1}^{q} \lambda_j \beta_j' \mathbf{K}_j \beta_j} \tag{10}$$

by means of Penalized Iterative Re-weighted Least Squares (P-IRLS) [7, 25]. We utilize the mixed model representation of penalized splines to determine the smoothing parameters based on restricted by maximum likelihood (REML).

The main advantage of the two-stage stepwise approach over classical stepwise regression is that it includes different modeling alternative per covariate and ensures that each covariate enters in only one specific form. The latter is particularly important to achieve an interpretable model when the algorithm terminates.

3.2 Double Penalty GAMs

Based on the same penalized likelihood scheme utilized in the two-stage stepwise approach, an approach for performing model selection and fitting in a single step can be developed by adding

another penalty for the selection of effects [17]. Given the equivalence of piecewise exponential models and Poisson model, this method is directly applicable once the required data augmentation of the original survival data has been performed. As mentioned before, the trade-off between fidelity to the data and smoothness of the estimated effects is governed by the penalty term $\lambda_j \beta_j' \mathbf{K}_j \beta_j$. However, even if there is no effect of the continuous covariate x_j, i.e. in cases where λ_j goes to infinity, there is no guarantee that that the corresponding smooth term is completely removed from the model. This is due to the fact that most penalties include a nontrivial null space consisting for example of linear functions in case of the difference penalties we are considering.

Hence, to have the possibility of shrinking the whole spline term to zero and therefore perform automatic model selection for the continuous variables, the inclusion of a second penalty term in the specification of (10)is proposed in [17]. Specifically, with this approach, each penalty term corresponding to a smooth term is replaced by

$$\boldsymbol{\lambda_j \beta_j' \mathbf{K}_j \beta_j + \lambda_j^* \beta_j' \mathbf{K}_j^* \beta_j} \tag{11}$$

where the first term in (11) penalizes functions on the range space and the second term penalizes functions in the null space. Note that categorical variables are not affected by this automatic model selection process.

3.3 Boosting

In a functional gradient descent boosting approach for estimating model (8), the estimation problem is reinterpreted as as the empirical risk minimization problem

$$\eta^*(\mathbf{z}) = \text{argmin}_{\eta(\mathbf{z})} \frac{1}{n'} \sum_{i=1}^{n'} \rho(y_i, \eta(\mathbf{z}_i)), \tag{12}$$

where ρ is a loss function defined by the Poisson negative log-likelihood

$$\rho(y_i, \eta(\mathbf{z}_i)) = -y_i \eta(\mathbf{z_i}) + \exp(\eta(\mathbf{z}_i)). \tag{13}$$

Minimization is then achieved by iteratively fitting suitable base-learning procedures to the negative gradients of the loss function that represent a natural measure for the lack of fit of the current model for a particular observation. Since we are not only interested in obtaining an estimate for the complete predictor (as it would be sufficient in prediction-oriented approaches) but mainly in model choice and variable selection, we use a componentwise boosting algorithm. Therefore, we specify separate base-learners g_j for each possible model alternative resulting in as many base-learners as model possibilities.

In our specific case, base-learners g_j are closely related to the functions f_j defined in (9) and we also use penalized splines to specify nonparametric base-learners for the smooth effects. In case of

parametric effects, simple least squares fits are used for the base-learner. In both cases, the base-learning procedure can be characterized by a corresponding hat matrix arising from a penalized least squares fit for the negative gradients, i.e.

$$\mathbf{Z}_j(\mathbf{Z}_j'\mathbf{Z}_j + \lambda\mathbf{K}_j)^{-1}\mathbf{Z}_j'$$

where $\lambda = 0$ in case of linear effects. To improve the model building properties of the boosting approach, we decompose each nonparametric effect in two base-learners following Kneib et al. (2009) [15]: one which contains the linear effect and another one which contains the nonparametric deviation from it. In case of boosting, the smoothing parameters λ_j are not estimated from the data but fixed before-hand since they only define the complexity of the base-learner while the complexity of the estimated functions is determined implicitly in the boosting algorithm. Assigning exactly one degree of freedom to each base-learner leads to an objective choice for the smoothing parameters of the base-learners and makes all base-learners comparable in terms of their complexity (Hofner et al 2011).

Functional gradient descent boosting then proceeds as follows:

(i) Initialize model all model components with $\widehat{f}_j^{[0]}(.) \equiv 0$, $j = 1, \ldots, J$; $\widehat{\eta}_i^{[0]}(.) \equiv \Delta_i$, $\rho_i^{[0]}(y_i, .) \equiv -y_i\eta_i^{[0]}(.) + \exp(\eta_i^{[0]}(.))$, $i = 1, \ldots, n'$. Set the iteration index to $m = 0$

(ii) Increase m by 1 and compute the negative gradients of the loss function evaluated at the current model fit:

$$u_i = -\frac{\partial}{\partial\eta}\rho(y_i, \eta)\Big|_{\eta=\widehat{\eta}^{m-1}(\mathbf{z}_i)}, i = 1, \ldots, n' \tag{14}$$

(iii) Determine the best-fitting componentwise base-learner $\widehat{g}_{j^*}$ according to the Poisson log-likelihood loss, i.e.:

$$j^* = \text{argmin}_{1\leq j\leq r}\sum_{i=1}^{n}(-y_i + \exp(\eta(\mathbf{z}_i)) - \widehat{g}_j(\mathbf{z}_i))^2. \tag{15}$$

(iv) Update the corresponding function estimate $\widehat{f}_j$ to

$$\widehat{f}_{j^*}^{[m]}(.) = \widehat{f}_{j^*}^{[m-1]}(.) + \nu g_{j^*}^{[m]}(.)$$

where $\nu \in (0, 1]$ is a step size. Remain all other effects constant and update $\widehat{\eta}_i^{[m]}(.) = \widehat{\eta}_i^{[m-1]}(.) + \nu\widehat{g}_{ij^*}^{[m]}(.)$ and $\widehat{\rho}_i^{[m]}(y_i, .) = -y_i\widehat{\eta}_i^{[m]}(.) + \exp(\widehat{\eta}_i^{[m]}(.))$.

(v) Store the reduction in the empirical risk achieved in the m-the step as

$$D^{[m]} = \sum_{i=1}^{n'}\widehat{\rho}_i^{[m-1]}(y_i, .) - \sum_{i=1}^{n'}\widehat{\rho}_i^{[m]}(y_i, .).$$

Steps (i) to (v) are iterated until $m = m_{stop}$. In practice, m_{stop} acts like a tuning parameter in the algorithm and is determined by resampling techniques such as bootstrapping or cross-validation.

A suitable, data-driven choice of the number of boosting iterations m_{stop} in combination with componentwise updates of only the best-fitting base-learner induces that some base-learners may never be selected and, hence, the boosting algorithm provides a direct measure of both variable and model selection. Moreover, the selection of a particular base-learner in step (iii) of iteration m induces a certain diminution on the empirical risk given by $D^{[m]}$. Hence, it is noticeable that not all the base-learners chosen in the boosting algorithm are equally important in terms of their contribution to the overall empirical risk reduction. Specifically, base-learners chosen in last iterations contribute usually have a smaller contribution to the overall risk reduction than base-learners chosen in earlier iterations. This specific characteristic of the boosting algorithm allows us to define a measure of relative importance of each model component in terms of the contribution to the overall the empirical risk reduction:

$$C(g_j) = \sum_{m \in I_j} D^{[m]} \tag{16}$$

where

$$I_j = \left\{ m \in 1, \ldots, m_{stop} \,\middle|\, j = \text{argmin}_{1 \leq j \leq r} \sum_{i=1}^{n} (-y_i + \exp(\widehat{\eta}_i^{[m-1]}(.)) - \widehat{g}_j(.))^2 \right\}.$$

3.4 Software Implementation

For practical application of the methods presented above, we implemented an easy to use function in R. This function requires the user to introduce the survival data set in their standard presentation consisting of the observed survival times, censoring indicator and vector of covariates and it conducts data augmentation in the manner explained in Section 2.2. The three aforementioned model building procedures are implemented using existing R packages. Specifically, package `mgcv` [25] was used to fit GAM models evolved in the two-stage stepwise and double penalty methods. In the latter case, the automatic selection of smooth terms is carried out via the argument `select=T` in function `gam`. The boosting algorithm is carried out by using the implementation of component-wise boosting in package `mboost` [13].

4 Simulation Study

4.1 Simulation Setup

To assess the performance of the piecewise exponential representation proposed in Section 2 in terms of flexibility and to compare the three aforementioned strategies for model building based on it, we conducted a simulation study. Specifically, we simulated 200 Monte Carlo trials based on a theoretical model in which both proportional hazards and log-linearity of covariate effects do not hold:

$$\lambda(t; x) = \exp\left[g_0(t) + \sum_{j=1}^{9} f_j(t, x_j)\right]$$

with t the observed survival time and $\mathbf{x} = (x_1, \ldots, x_9)'$ the vector of informative covariates. $x_1, \ldots, x_9$ were sampled as follows: x_1 to x_4 are continuous while x_5 to x_9 are binary. Specifically $x_1, x_2 \overset{\text{i.i.d.}}{\sim} U(-1, 1)$, $x_3, x_4 \overset{\text{i.i.d.}}{\sim} N(0, 1)$, $x_5, x_9 \overset{\text{i.i.d.}}{\sim} B(1, 0.5)$, and x_6, x_7, x_8 are correlated binary covariates ($\phi(x_i, x_j) = 0.75$; $i, j = 6, 7, 8$). They are the result of the categorization (using the mean as cut-off point) of three variables sampled from a multivariate normal distribution $N_3(\mu, \Sigma)$, where $\mu = (0, 0, 0)'$ and variances in Σ are defined by a parameter $\rho = 0.9$ which governs the level of correlation among covariates. The functions defining the linear predictor are listed in Table 1.

Table 1: Simulation study. Specification of the theoretical linear predictor.

$g_0(t)$=2+log(t+0.2)
$f_1(x_1) = \sin(-x_1^2 - 0.6x_1^3)$
$f_2(x_2) = -0.3x_2$
$f_3(x_3) = 0.5\sin(1.5x_3)$
$f_4(x_4) = x_4$
$f_5(x_5) = -2x_5$
$f_6(x_6) = 0.15x_6$
$f_7(x_7) = 0.20x_7$
$f_8(x_8) = 0.25x_8$
$f_{t,9}(t, x_9) = 2\sqrt{t}x_9$

Theoretical survival times T_i are generated by means of an inversion technique proposed to simulate survival times from Cox-type models [1]. This method allows the simulation of Cox-type models including time-varying effects with arbitrary baseline hazards as long as the cumulative hazard and its inverse are available, at least for numerical evaluation. To obtain censored observations, we generated $C_i \sim Exp(c)$, where c is a constant which controls the amount of censoring and we defined the observed survival time as $t_i = \min(T_i, C_i)$. With regard to evaluate the variable selection

properties of the implemented methods, we also generated six nuisance covariates without association with the hazard function. Hence the covariate vector is composed of 15 covariates: $\mathbf{x} = (x_1, \ldots, x_{15})'$ where $x_1, \ldots, x_9$ are the aforementioned nine informative covariates and $x_{10}, \ldots, x_{15}$ are nuisance covariates. Specifically, $x_{10}, \ldots, x_{12}$ are continuous correlated covariates sampled from a multivariate normal distribution and $x_{13}, \ldots, x_{15}$ are binary independently draw from $B(1, 0.5)$.

As basic simulation setting to illustrate our findings, we considered $n = 250$ as sample size and 40% of censored observations. We also checked how enlarging sample size to $n = 500$ and increasing the number of nuisance variables affect simulation results. Main results about these complementary scenarios are also discussed in the text, but detailed results are not displayed for sake of brevity. Additional plots and tables concerning them are given as supplemental material.

With regard to the data augmentation strategy, we assumed a grid of equidistant jump points along the range of variation of t. Specifically, we considered 25 points in our basic simulation setting. As a result, the median length of the augmented samples for each trial was $n' = 1240$ when using an original sample size $n = 250$.

The goodness of fit was measured in terms of the empirical mean squared error (MSE) in estimating linear predictors and theoretical effects of covariates, i.e.

$$MSE(\widehat{f}_j) = \frac{1}{n'} \sum_{i=1}^{n'} (f_j(x_i) - \widehat{f}_j(x_i))^2, i = 1, \ldots, n'$$

For binary variables effects, the MSE is computed as the squared difference between theoretical and estimated coefficients.

To evaluate model building performance, the empirical norm of effects for both informative and nuisance variables was calculated.

$$||\widehat{f}_j|| = \sqrt{\sum_{i=1}^{n'} (f_j(x_i))^2}, i = 1, \ldots, n'$$

Since the non relevance of an effect estimate can be defined as its closeness to zero, we expect that the spurious effects present a norm equal to zero. We found the norm preferable to other possibilities, as for example the number of times that a given nuisance variable or an incorrect model term is chosen, since it allows to evaluate not only the wrong choice but its magnitude and makes the three methods comparable. This is specially important for both the one-step methods studied: in the case of boosting it is noteworthy that the selection of a particular base-learner has not the same impact on its contribution if it has been produced at the beginning or at the end of the iterative process. Moreover, the double penalty approach does not allow to remove completely categorical covariates but the magnitude of the assigned effect is important to measure its effective model building capacity.

4.2 Simulation Results

In terms of goodness of fit, Figure 1 shows the theoretical smooth effects and the corresponding mean estimated effect along the 200 trials for the three methods discussed here, whereas Table 2 shows the median MSE and interquartile range for estimating the linear predictor and effects of informative covariates. All function estimates and true effects were centered so that their mean is equal to zero to avoid displacements due to differences in level of estimations in each method.

Variable selection and model choice performance is visualized in Figure 2. The median, and first and third quartiles of the empirical norms of each effect included in the models were used as summary statistics and graphically represented. On the top panel, we present results for the main effects. We expect the empirical norm to be zero for variables $x_{10}, \ldots, x_{15}$. In the bottom panel, the time-varying effects included in the models are represented in the same way. Models with a good model choice performance should assign norm zero to all time-varying effects with the exception to the corresponding to x_9, truly informative.

From Figure 1 we observe that the piecewise exponential representation allows to capture the smooth shape of both baseline hazard and the time-varying effect of x_9, specially for the two-stage stepwise and double penalty methods. Boosting presents a poorer performance for estimating the true log-baseline hazard, shrinking the true effect towards zero, however this conservative performance of boosting is common to the rest of the continuous covariates, and hence not particularly related to the piecewise exponential representation.

The estimated median MSE of the overall linear predictor is only slightly different across the three compared approaches. However, the bad performance of the two-stage stepwise method when estimating the effect of continuous covariate x_4 with a linear true effect and the binary covariate x_5 is noteworthy. The explanation is found in Figure 2 where we can observe how this method wrongly assigns false time-varying effects to these covariates. It is noteworthy that even when increasing the sample size, the biased selection towards spurious time-varying effects of these two truly parametric effects of the two-stage stepwise method persists. It seems that this tendency towards overcomplicated estimation of true parametric effects is also related to the magnitude of the theoretical effect, affecting specially the most relevant covariates.

If we focus in the continuous covariate x_2 which is also associated to a linear but weaker true effect, the corresponding MSE is much lower, however, we can notice how the two-stage stepwise method also presents symptoms of the same biased selection towards time-varying effects (Figure 2 bottom panel shows a large third quartile for its corresponding time-varying effect).

With regard to informative binary and correlated covariates $x_6, \ldots, x_8$, both two-stage stepwise and

boosting tend to underestimate the true effects whereas the double penalty approach tends to the overestimation and higher variability in the estimates.

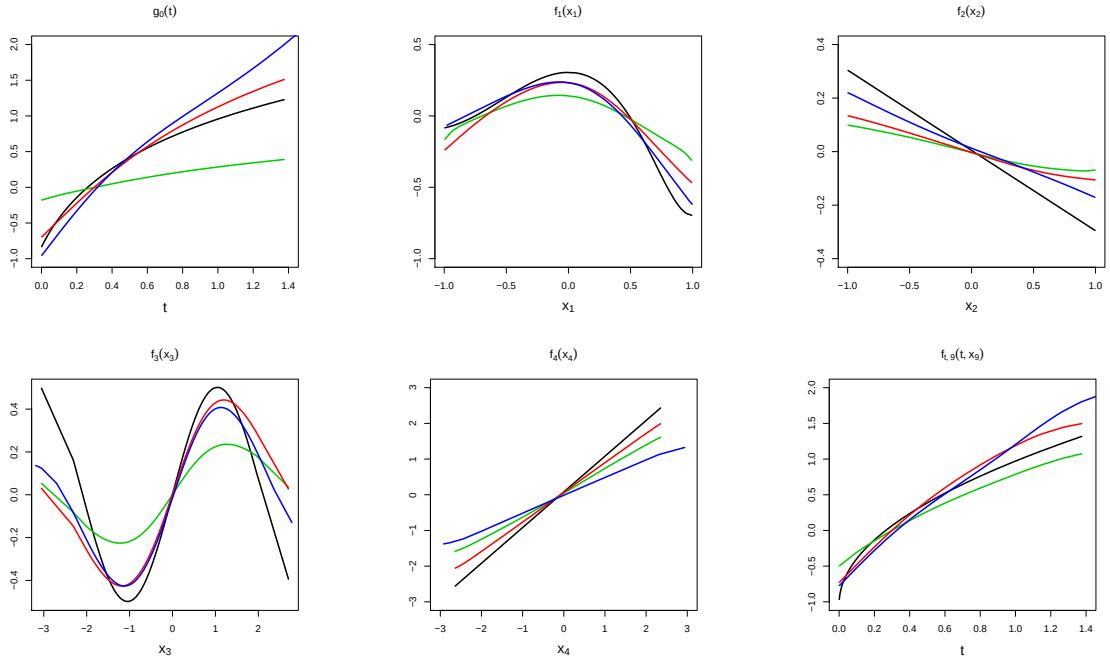

Figure 1: Simulation study. Estimated mean effects along the 200 trials of the continuous covariates included in the theoretical linear predictor. The black line represent the true effects and the colored ones the estimated effects. Blue lines: Two-stage stepwise method. Red lines: Double penalty method. Green lines: Boosting method.

Table 2: Simulation study. Estimated median MSE and interquartile range (IQR) along 200 trials of the three model building methods. $n = 250$; 40% censoring.

Theoretical quantity	Two-Stage stepwise	Double Penalty GAM	Boosting
η	2.699(1.219)	2.522(0.995)	2.453(0.806)
g_0	0.160(0.313)	0.083(0.119)	0.220(0.215)
f_1	0.024(0.043)	0.031(0.029)	0.042(0.036)
f_2	0.028(0.024)	0.028(0.019)	0.026(0.021)
f_3	0.037(0.042)	0.026(0.026)	0.045(0.045)
f_4	0.430(0.965)	0.012(0.049)	0.106(0.105)
f_5	4.000(3.917)	0.032(0.112)	0.295(0.417)
f_6	0.022(0.001)	0.041(0.134)	0.022(0.001)
f_7	0.040(0.000)	0.049(0.183)	0.040(0.028)
f_8	0.062(0.000)	0.051(0.126)	0.044(0.054)
$f_{t,9}$	0.068(0.111)	0.047(0.065)	0.059(0.083)

In terms of variable selection and correct model choice, Figure 2 gives interesting information. From top panel, we can observe that both two-stage stepwise and boosting methods are able to discard spurious variables $x_{10}, \ldots, x_{15}$. The double penalty approach also presents a good performance with regard to dropping spurious continuous covariates $x_{10}, \ldots, x_{12}$, but this method presents problems to shrink the effect of spurious binary covariates towards zero. Due to the semi-automatic nature of

this method, in which categorical covariates can not be automatically removed, we observe that it overestimates the effect of binary variables included in the model, both the informative and the not relevant ones.

From Figure 2, bottom panel, we observe that the two-stage stepwise method is more able to completely remove spurious time-varying effects corresponding to non informative covariates than double penalty and boosting approaches. However, as mentioned, the performance of two-stage stepwise method is dramatically bad when referring to correct modeling of true parametric effects. On the other hand, boosting and double penalty GAM present a similar performance in relation to the selection of spurious time-varying effects. Although the median empirical norm corresponding to these spurious terms is zero when using either of them, the third quartile estimates indicates some tendency to false positive time-varying effects discoveries, specially when using the double penalty method. Even if this is a not desirable property, we nicely observe, that both methods are able to distinguish in terms of empirical norm magnitude, among true and false time-varying effects (third quartile of any spurious time-varying estimate is lower than the first quartile of the true time-varying effect corresponding to variable x_9).

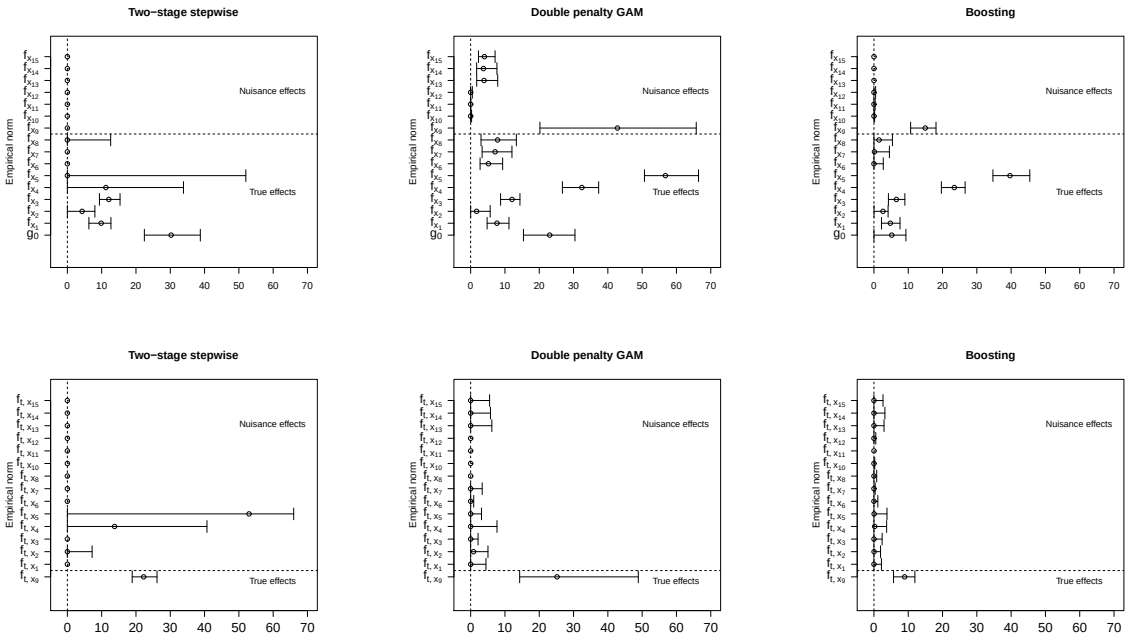

Figure 2: Simulation study. Comparison of model choice performance among methods. The median of the empirical norm along the 200 trials and the first and third quartile are represented for each of the model components included in the three methods. Top panel: Main effects. Bottom panel: Time-varying effects. In each plot, a dashed line separates truly informative components from the nuisance components.

When moving to a scenario with more nuisance covariates (from 6 to 21), the conclusions remain basically the same. For the boosting it seems more difficult to capture the sinusoidal effect of

covariate x_3 than in the basic setting, while is more stable than the other approaches to discarding both spurious covariates and time-varying effects.

5 Application: Prognosis After Myocardial Infarction

5.1 Data Presentation and Motivation

In this section, we aim to get insights in factors affecting long-term survival for patients who have suffered a myocardial infarction. Myocardial infarction (MI) is a disease in which the development of more precise estimates of risk and prognosis is desirable, as it can result in serious and fatal outcomes. Moreover, efficacious therapies for this disease have been developed during the last two decades. This further amplifies the need for prognostic prediction on which to form an understanding of future expectations and to base therapeutic and other management decisions so as to reduce the associated short and long-term morbidity and mortality. With this application, we aim to investigate which are the most relevant predictors of survival after myocardial infarction and their optimal degree of complexity considering a high number of predictors simultaneously. The study subjects were all patients admitted consecutively between 1/2006 and 2/2011 to the Cardiology department of the Complexo Hospitalario Universitario de Santiago de Compostela, Galicia, Spain and having a final diagnosis of myocardial infarction (n=3.027). We define as event of interest 1-year mortality and as potential variables affecting prognosis, we considered 26 variables referring to demographic, clinical and angiographic characteristics, as well as variables related to in-hospital management and complications. They are listed and shortly described in Table 3. Despite most of them are self-explanatory and widely known prognostic factors for acute coronary syndrome, we also consider less studied factors, such as the effect of in-hospital bleeding events or the ST-segment deviation observed in the admission electrocardiogram or the possibly protective effect of performing exploratory procedures such as a coronary angiography. Specially the role of hemorrhagic complications in prognosis after a myocardial infarction is a question of recent concern and still subject to debate. Moreover, it does not exist a consensus about what is the best definition of a hemorrhagic complication, i.e. different criteria exist for measuring it. Here, we considered two of the most accepted bleeding scales, the classical TIMI bleeding criterion (Thrombolysis in Myocardial Infarction) [5] and the most recently proposed BARC bleeding criterion (Bleeding Academic Research Consortium)[18]. Both of them are highly related since they are derived from overlapping definitions, being TIMI more restrictive than BARC. Hence, we can simultaneously address the problem of determining if bleeding is important and which is the best scale to measure it. Note that the aim

of our application is to identify and establish hypothesis but not to make individual predictions for each patient.

Table 3: Application: Prognostic after myocardial infarction. Definition of the covariates included in the models as potentially related to prognosis.

Covariate	Type	Description
Demographic		
Age	Continuous	Age (in years) at admission
Sex	Binary	Sex (Female=1)
Past medical history		
PAD	Binary	Peripheral artery disease
CHF	Binary	Congestive heart failure
Malignancy	Binary	Malignancy
AF	Binary	Chronic atrial fibrilation
CAD	Binary	Coronary artery disease
DM	Binary	Diabetes mellitus
On-admission data		
ST	Binary	ST segment deviation
Anemia	Binary	Anemia (WHO criteria)
HT	Binary	Hypertension
IAM-EST	Binary	Non ST-elevation myocardial infarction
Killip	Categorical(4 values)	Killip class (I-IV)
HR	Continuous	Heart rate (beat/minute)
Hb	Continuous	Hemoglobin concentration(g/dL)
cTnI	Continuous	Cardiac Troponine I levels(ng/dL)
log(WBC)	Continuous	logarithm of white cells count(giga/L)
eRF	Continuous	Estimated renal function (Cockcroft-Gault formula)
In-hospital management		
PCI	Binary	Percutaneous coronary intervention
CATH	Binary	Coronary angiography
TIMI	Binary	Hemorrhagic complications (TIMI scale)
BARC	Binary	Hemorrhagic complications (BARC scale)
Treatment after discharge		
B-blocker	Binary	Treatment with B-Blockers
Statin	Binary	Treatment with Statin
Aldosterone	Binary	Treatment with Aldosterone Blockers
ACEARB	Binary	Treatment with ACE or ARB inhibitors

5.2 Results

Table 4 shows the estimated effects for the categorical variables included in the study according to the three analyzed model building strategies and their relative importance in terms of the empirical norm. For sake of better interpretability, we express these quantities in percentages with respect to the total contribution, defined as the sum of the empirical norms of all included terms in each corresponding model. For the boosting method, we also provide the specific contribution measure

based on empirical risk reduction introduced in Section 3.3. given as percentage with respect to the total empirical risk reduction.

From Table 4, we observe that both two-stage stepwise and boosting methods show a high level of agreement with regard to categorical variables selection: anemia, Killip class, previous malignancy, heart failure and cardiac disease are identified as the most important risk factors at baseline whereas treatment with B-blockers and the performance of a percutaneous coronary intervention (PCI) as the most relevant protective factors among categorical covariates. Additionally, boosting method also chose coronary angiography (CATH) as a protective factor. In agreement with results from simulation, the size of the effects provided by boosting method are lower than the corresponding to two-stage stepwise method.

Despite the double penalty GAM method does not carry out an effective variable selection of the parametric effects of categorical variables, the strongest estimated effects with this method correspond, in general, to those variables automatically chosen with the other proposals.

Moreover, the double penalty approach provides reasonable, although probably overestimated, according to simulation results, baseline effects for previously reported risk factors as hypertension (HT), previous peripheral arterial disease (PAD) and ST-segment deviation (ST) and protective factors as feminine sex, treatment with with ACE or ARB inhibitors and non ST-elevation myocardial infarction (NSTEMI).

However, the double penalty approach also provides not clinically plausible baseline effects of some categorical covariates. Namely, the estimated protective effect of chronic atrial fibrilation (AF) and the identification of the treatment with Aldosterone Blockers as a risk factor are hardly clinically interpretable. Special mention requires the modeling of the four category covariate Killip class. This covariate reflects an increasing severity of the heart failure from the mildest damage represented by category II to the most dramatic cardiogenic shock represented by category IV. Killip class I refers to individuals with no clinical signs of heart failure and is considered as reference. According to that, one would expect an increasing effect along Killip categories. All the three methods assign the strongest effect to Killip III category. Given the limited number of patients with Killip IV (n=40) in our sample, it is not surprising to find a similar effect for Killip II and IV categories, as the boosting method indicates. However, both double penalty and two-stage stepwise methods assign a non clinically plausible baseline protective effect to Killip IV category. Specifically, the double penalty approach provides an overcomplicated model proposal for Killip IV class: a combination of a strong baseline protective effect which is corrected by an increasing time-varying effect.

With regard to the modeling of the two dummy covariates accounting for bleeding complications

(TIMI and BARC), the three analyzed methods also differ. The two-stage stepwise method does not include any of these covariates in the final model, while the boosting method chooses a pure time-varying effect to TIMI scale, removing completely BARC. However, the double penalty approach assigns a strong risk effect of BARC jointly with a protective baseline effect of TIMI complemented by a weak time-varying effect. Again, it seems a too complicated and hard to interpret model proposal.

Table 4: Application: Prognostic after myocardial infarction. Estimated effects of categorical covariates and relative contribution in terms of the relative norm (||.||(%))and boosting specific empirical risk-reduction measure (C(%)).

Covariate	Two-Stage stepwise		Double Penalty		Boosting		
	Effect	||.||(%)	Effect	||.||(%)	Effect	||.||(%)	C(%)
Malignancy	0.72	6.38	0.65	5.30	0.42	8.64	1.74
CHF	0.56	4.69	0.43	3.01	0.38	7.41	35.36
Anemia	0.53	5.19	0.53	4.31	0.40	8.93	4.84
PCI	-0.47	6.45	-0.41	4.67	-0.17	5.28	0.64
CAD	0.41	4.16	0.36	3.03	0.20	4.59	0.66
B-blocker	-0.41	5.73	-0.34	3.95	-0.23	7.48	1.26
Killip II	0.25		0.27		0.41		
Killip III	0.72	7.61	0.80	26.99	0.59	16.21	12.47
Killip IV	-0.50		-4		0.40		
CATH	0.00	0.00	-0.28	3.61	-0.18	6.59	12.22
Sex (Woman)	0.00	0.00	-0.34	3.11	0.00	0.00	0.00
TIMI	0.00	0.00	-0.10	0.68	0.00	0.00	0.00
BARC	0.00	0.00	0.50	3.51	0.00	0.00	0.00
NSTEMI	0.00	0.00	-0.32	0.83	0.00	0.00	0.00
HT	0.00	0.00	0.11	1.21	0.00	0.00	0.00
AF	0.00	0.00	-0.17	1.22	0.00	0.00	0.00
DM	0.00	0.00	-0.08	0.57	0.00	0.00	0.00
PAD	0.00	0.00	0.11	0.82	0.00	0.00	0.00
ST	0.00	0.00	0.11	1.16	0.00	0.00	0.00
Statin	0.00	0.00	-0.09	4.28	0.00	0.00	0.00
Aldosterone	0.00	0.00	0.20	1.43	0.00	0.00	0.00
ACEARB	0.00	0.00	-0.19	2.12	0.00	0.00	0.00

Baseline hazard and smooth main effects (when selected by at least one of the methods) of continuous covariates are presented in Figure 2. Cardiac Troponin I levels (cTnI) is not chosen by none of the methods, whereas Hemoglobin is only chosen by the double penalty approach, resulting in a non clinical plausible estimate according to previous bibliography [21]. With regard to age, the three methods identify it as an important risk factor, however, they disagree in terms of the assigned complexity. Within the two-stage stepwise approach, age is modeled as a time-varying effect, specifically, this method assigns a linear and decreasing effect along the follow-up time to age. The corresponding estimate to $t = 0$ of follow-up time is represented in Figure 2. On the other

hand, both other methods agree in assigning a fixed-time effect to age. Moreover, while the double penalty assigns an increasing effect close to linearity along the range of age, boosting only identifies as risky the ages from 60 years old. For older patients, both boosting and double penalty method estimates of age almost coincide. A high level of agreement among the three methods is observed for both heart rate and estimated renal function. The main difference across methods in terms of continuous covariates was found for the model choice of the logarithm of white blood cells count (log(WBC)). Boosting shows a weak and slightly increasing effect while the two GAM methods identify a markedly U-shaped effect hardly clinically interpretable [19].

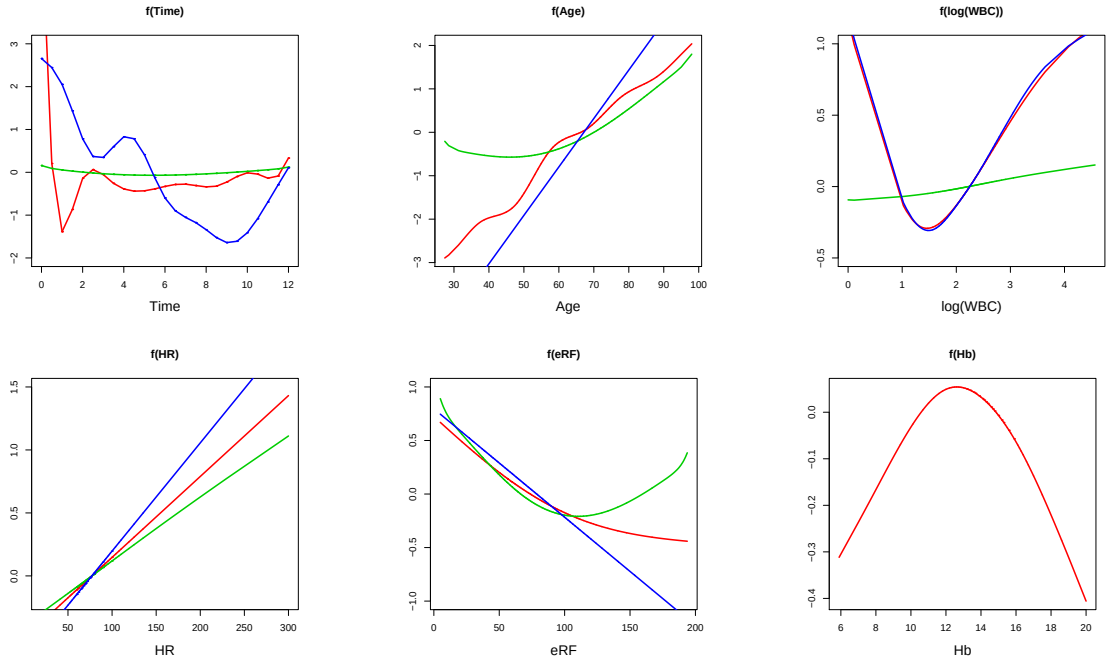

Figure 3: Application: Prognostic after myocardial infarction. Estimated effects (when chosen) of the continuous covariates by the three methods. cTnI is never chosen for none of the methods. Blue lines: Two-stage stepwise method. Red lines: Double penalty method. Green lines: Boosting method. The estimated effect of age by the two-stage stepwise method is time-varying, the presented effect corresponds to $t = 0$ of follow-up.

A summary of the time-varying effects chosen by at least one method and their relative contributions is given in Table 5. We observe that, apart from age, the two-stage stepwise method does not select any other time-varying effect as relevant. However, both double penalty and boosting methods do not completely remove a considerably number of time-varying effects. The double penalty approach chose seven time-varying effects, but only the corresponding to ST-segment deviation and Killip IV seem relevant, with regard to their relative contribution. These two terms are associated to a relative empirical norm of 1.5% or higher, whereas the rest of the terms present a relative contribution of 0.72% at the most. As for the boosting method, it chose 12 time-varying terms in at least one iteration until the termination of the algorithm, but their relative contributions seem negligible

except for TIMI bleeding, ST-segment deviation and hypertension, with contributions larger than 2.5% in terms of relative empirical norm and larger than 1.3% when using the empirical risk reduction contribution measure C. The remaining time-varying effects are associated to a relative contribution lower than 1% in terms of empirical norm and lower than 0.5% in terms of C. Despite different in scale, it is interesting to observe how the new specific measure of contribution derived for boosting algorithms yields to the same conclusions than those derived from the empirical norm. The time-varying effects for TIMI, ST-segment deviation and hypertension are represented in Figure 3. ST-segment deviation present a decreasing with time effect, whereas TIMI bleeding and hypertension present an approximately spoon shape. From these results, it seems feasible and useful to fix a cut-off value in terms of relative contribution to neglect the less relevant time-varying effects and fix a more easy to use final model. In this application, we propose to reject time-varying effects which represent less than the 1% in terms of relative contribution, but more restrictive conditions can be imposed in practice, for both time-varying and main effects.

Table 5: Application: Prognostic after myocardial infarction. Estimated time-varying effects (when chosen by at least one method) and relative contribution in terms of the relative norm (||.||(%))and boosting specific empirical risk-reduction measure C(%)

Covariate	Two-Stage	Double Penalty	Boosting	
	||.||(%)	||.||(%)	||.||(%)	C(%)
ST	0.00	1.60	2.95	1.47
TIMI	0.00	0.48	2.81	1.33
HT	0.00	0.00	5.36	2.61
Killip (IV)	0.00	2.89	0.90	0.46
Anemia	0.00	0.00	0.78	0.42
CHF	0.00	0.13	0.70	0.33
AF	0.00	0.00	0.43	0.13
Age	35.74	0.00	0.18	0.06
CAD	0.00	0.00	0.86	0.30
Malignancy	0.00	0.00	0.29	0.07
PAD	0.00	0.00	0.16	0.03
IAM-EST	0.00	0.72	0.00	0.00
DM	0.00	0.28	0.36	0.13
cTnI	0.00	0.42	0.00	0.00
B-blocker	0.00	0.59	0.00	0.00
Statin	0.00	0.19	0.00	0.00
PCI	0.00	0.35	0.00	0.00

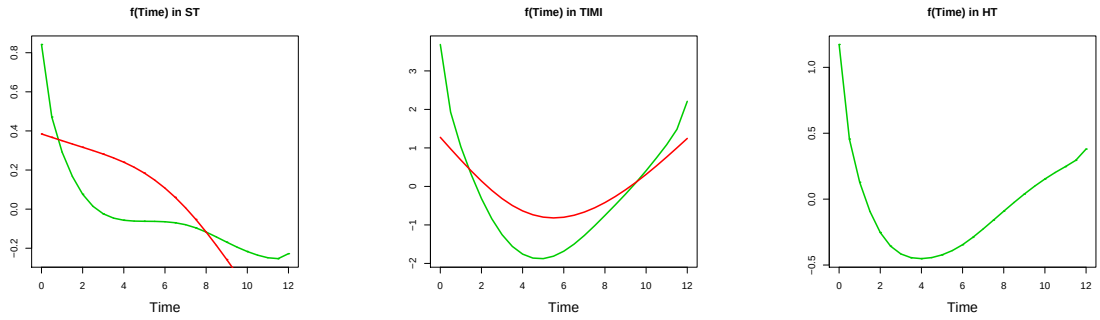

Figure 4: Application: Prognostic after myocardial infarction. Estimated time-varying effects (when chosen) of ST segment deviation (ST), TIMI bleeding and hypertension (HT). Red lines: Double penalty method. Green lines: Boosting method.

From this application, we found that all the three analyzed methods provided similar and clinically plausible results with regard to which presumably are the most strongly related to survival after myocardial infarction factors. However, both two-stage stepwise and the double penalty approaches show tendency to overfitting. Specially, the double penalty approach produce annoying modeling artifacts for some specific covariates such as Killip class and bleeding complications which make this method less suitable than the others in this application. Boosting seems to be more overfitting resistant although it underestimates main effects and it detects a high number of time-varying effects. However, most of them could be in practice neglected according to their low relative contribution. With respect to time-varying effects selection, the two-stage stepwise procedure presents the advantage of not choosing low contribution terms, which yields to a more interpretable model but on the other hand it assigns a possibly overfitted time-varying effect to age. Altogether, all the three approaches present advantages and drawbacks and the preferable one or a combination of methods should be given by the specific practical situation.

6 Summary and Discussion

In this paper, we proposed and compared several strategies for model building in non proportional hazard regression based on a piecewise representation of the original survival data. We focused on the situation where model building is the ultimate goal of the analysis, i.e., when the problem to be solved is to choose a subset of covariates and their corresponding modeling alternatives for given survival data. Although estimation for complex multivariate models is relatively straightforward using available software, selection of the 'best' prediction model is a much more subtle task, specially in the survival analysis framework.

We revisited the equivalence in terms of likelihood between piecewise exponential representation of survival data and count data by means of data augmentation of the original dataset. Such

transformation allowed us to link hazard regression with estimation schemes based on the Poisson likelihood and to adapt recent advances for model building in exponential family regression to the non proportional hazard regression context.

Specifically, we focused in three proposals coming from different conceptual frameworks: a two-stage stepwise selection procedure, a shrinkage approach based on double penalization likelihood and a componentwise functional gradient descent algorithm were adapted and compared in this context. All them consider a set of candidates terms to be included in a general additive predictor, are based on Poisson likelihood, and use penalized splines as smoothers.

We gained valuable insights into the performance and the practical usability of all three approaches under investigation by means of simulation and an application to a real biomedical dataset. Simulations show that for finite samples, the piecewise exponential representation shows a reasonably accurate performance in estimating smooth baseline rates and it can be seen as a nonparametric modeling approach when taking the intervals small enough .

Our results show that clinically plausible prognostic models can be constructed with the proposed methods, but some notes of caution should be given. With regard to the two-stage stepwise procedure, both simulation and real data analysis results suggest that this method presents a bias selection towards more complicated effects than required when a covariate is truly informative. However, it has the advantage of choosing a unique model alternative for each covariate and the ability of removing completely time-varying effects of non informative covariates which makes the resulting model appealing for interpretation. The predictive criterion AIC may be behind its tendency to overfitting. Focusing on a different criterion may improve the overall performance of this method. On the other hand, the semiautomatic double penalty GAM approach presents the drawback that it does not allow for the automatic choice of categorical effects, which results in an inflated model, with more covariates than necessary. Moreover, as it has been observed in the real data application, it can lead to overcomplicated combination of effects, which are not transportable at all to clinical practice. The main characteristic of boosting observed in simulation and enforced with the application is the tendency to shrink effects towards zero. With regard to false time-varying effects selection analyzed by means of simulation, both boosting and double penalty are more prone to it than the two-stage stepwise procedure. However, when accounting for the contribution of these false discovery performance in terms of the empirical norm, it has been shown that these terms provide very low contribution and in practice could be neglected.

In order to neglect low contribution terms, in the real data application, we propose to take a pragmatic choice based on a minimum value of relative contribution of the not automatically removed

terms. We propose to fix this cut-off in terms of the empirical norm, or the reduction risk based contribution measure defined for boosting.

From a practical point of view, a good model building procedure should pursue parsimony to guarantee interpretability, stability and clinical usefulness. Keeping that in mind, a practitioner is interested in methods that are able to detect strong factors, strong deviations from linearity for continuous covariates and strong time-varying effects and that neglect covariates with a weak effect, wiggly functional forms of continuous covariates and weak deviations from proportionality of hazards [23]. Altogether, boosting seems to be the preferable method of the three studied alternatives. Despite it presents a tendency to shrink effects towards zero in comparison to the other proposals, its competitors, specially in the real data application, present some overfitted and not clinically plausible effects that are avoided by the boosting method. Since the aim of our application is explanatory, boosting seems to provide more interpretable effects and reproducible models in new data sets.

The Poisson approach to survival analysis presents other advantages that should be exploited in the future, among which the easy modeling of sophisticated risk patterns, the possibility to include time-varying covariates, etc. Moreover, all the proposals can be extended with the inclusion of interaction terms among specific covariates in complete analogy to the time-varying effects already considered.

From a more general and applied point of view, dealing with non proportional hazard regression by means of piecewise exponential models and data augmentation has the advantage that any methods proposed in the GAM field can be easily adapted to the survival context.

Penalized splines were used to represent the smooth effects of continuous covariates, including time and time-varying effects. Other possibilities for flexible modeling of effects could also be adapted to the presented framework, such as fractional polynomials [20, 22], can be alternatives to the penalized splines method used in this paper.

All the proposed methods were implemented in R and an easy to use function for conducting the methods presented in this paper is available. The code is available from the first authors upon request.

Acknowledgements

Work supported by the grants MTM2008-03129, MTM2011-23204, MTM2011-28285-C02-01 and MHE2011-00110 of the Spanish Ministerio de Ciencia e Innovacion and INBIOMED 2009-063 of the Xunta de Galicia. Support from BioStatNet: an interdisciplinary Biostatistics network (MTM2010-09213-E, MTM2011-15849-E) is also acknowledged.

References

[1] Bender, R., Augustin, T., and Blettner, M.(2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* **24**, 1713 – 1723.

[2] Binquet, C., Abrahamowicz, M., Mahboubi, A., et al. (2008). Empirical study of the dependence of the results of multivariable flexible survival analyses on model selection strategy. *Statistics in Medicine* **27(30)**, 6470 – 6488.

[3] Buchholz, A. and Sauerbrei, W. (2011). Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data. *Biometrical Journal* **53(2)**, 308 – 331.

[4] Bühlmann, P., and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* **22**, 477 – 505.

[5] Chesebro, J.H., Knatterud G., Roberts R., et al. (1987). Thrombolysis in Myocardial Infarction (TIMI) Trial, phase I: a comparison between intravenous tissue plasminogen activator and intravenous streptokinase: clinical findings through hospital discharge. *Circulation* **76**, 142 – 154.

[6] Cox, D.R.(1972). Regression models and life tables (with discussion). *J R Stat Soc Series B* **34**, 187 – 220.

[7] Eilers, PH. and Marx, BD.(1996). Flexible smoothing using B-splines and penalties. *Stat Sci* **11**, 89 – 121.

[8] Friedman M.(1982). Piecewise Exponential Models for Survival Data with Covariates. *The Annals of Statistics* **10**, 101 – 113.

[9] Hastie T. and Tibshirani R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B* **55(4)**, 757 – 796.

[10] Hennerfeind A., Brezger A., and Fahrmeir L. (2006). Geoadditive survival models. *J Am Stat Assoc* 101:1065-1075.

[11] Hofner, B., Kneib, T., Hartl, W. and Küchenhoff, H. (2011). Building Cox-type structured hazard regression models with timevarying effects. *Statistical Modelling* **11(1)**, 3 – 24.

[12] Holford, T. R. (1980). The analysis of rates and survivorship using log-linear models. *Biometrics* **36**, 299 – 305.

[13] Hothorn, T., Bühlmann, P., Kneib, T., et al. (2012). mboost: Model-Based Boosting, R package version 2.1-2. `http://CRAN.R-project.org/package=mboost`.

[14] Kneib T. and Fahrmeir L.(2007). A mixed model approach for geoadditive hazard regression. *Scand J Statist* 34: 207-228.

[15] Kneib, T., Hothorn, T., and Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics* **65(2)**, 626 – 634.

[16] Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques.. *Journal of the American Statistical Association* **76**, 231 – 240.

[17] Marra, G., Wood, S.N.(2011). Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis* **55**, 2372 – 2387.

[18] Mehran, R., Rao, S.V., Bhatt, D.L., et al. (2011). Standardized bleeding definitions for cardiovascular clinical trials. *Circulation* **123**, 2736 – 2747.

[19] Palmerini, T., Mehran, R., Dangas, G. et al. (2011). Impact of Leukocyte Count on Mortality and Bleeding in Patients With Myocardial Infarction Undergoing Primary Percutaneous Coronary Interventions. *Circulation* **123**, 2829 – 2837.

[20] Royston, P. and Altman, D.G.(1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with Discussion). *Applied Statistics* **43(3)**, 429 – 467.

[21] Sabatine, M.S., Morrow, D.A., Giugliano, R.P. et al. (2005). Association of Hemoglobin Levels With Clinical Outcomes in Acute Coronary Syndromes. *Circulation* **111**, 2042 – 2049.

[22] Sauerbrei, W. and Royston, P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society, Series A* **162**, 71 – 94.

[23] Sauerbrei, W., Royston, P. and Binder, H. (2007). Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine* **26**, 5512 – 5528.

[24] Tutz, G. and Binder, H.(2006). Generalized additive modelling with implicit variable selection by likelihood-based boosting. *Biometrics* **62**, 961 – 971.

[25] Wood, S.N. (2006). Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Appendix:
Supplemental results from simulation

Supplemental scenario 1. $n = 500$, 40% of censored observations. 15 covariates included in each model.

This scenario follows the same specifications than the one presented in the main text. In this case, sample size is enlarged to $n = 500$.

Specifically, the models include 15 covariates, 9 informative and 6 non-informative.

$x_1, \ldots, x_9$ are the informative covariates and were sampled as follows: x_1 to x_4 are continuous while x_5 to x_9 are binary. Specifically $x_1, x_2 \overset{\text{i.i.d.}}{\sim} U(-1, 1)$, $x_3, x_4 \overset{\text{i.i.d.}}{\sim} N(0, 1)$, $x_5, x_9 \overset{\text{i.i.d.}}{\sim} B(1, 0.5)$, and x_6, x_7, x_8 are correlated binary covariates ($\phi(x_i, x_j) = 0.75$; $i, j = 6, 7, 8$). They are the result of the categorization (using the mean as cut-off point) of three variables sampled from a multivariate normal distribution $N_3(\mu, \Sigma)$, where $\mu = (0, 0, 0)'$ and variances in Σ are defined by a parameter $\rho = 0.9$ which governs the level of correlation among covariates.

$x_{10}, \ldots, x_{15}$ are the nuisance covariates and were sampled as follows: $x_{10} - x_{12}$ are continuous correlated covariates sampled from a multivariate normal distribution and $x_{13} - x_{15}$ are binary independently draw from $B(1, 0.5)$.

Table 6: Estimated median MSE and interquartile range (IQR) along 200 trials of the three methods.$n = 500$, 40% censoring. 15 covariates (medium noise level)

Theoretical quantity	Two-Stage stepwise	Double Penalty GAM	Boosting
η	1.627(0.570)	1.597(0.567)	1.671(0.547)
g_0	0.084(0.129)	0.046(0.041)	0.110(0.128)
f_1	0.009(0.012)	0.017(0.022)	0.027(0.021)
f_2	0.006(0.027)	0.017(0.027)	0.011(0.022)
f_3	0.014(0.014)	0.016(0.015)	0.027(0.023)
f_4	0.910(0.961)	0.007(0.026)	0.036(0.046)
f_5	4.000(0.000)	0.021(0.059)	0.173(0.236)
f_6	0.022(0.000)	0.022(0.049)	0.022(0.017)
f_7	0.040(0.000)	0.017(0.061)	0.018(0.034)
f_8	0.062(0.029)	0.025(0.064)	0.024(0.051)
$f_{t,6}$	0.040(0.029)	0.032(0.023)	0.034(0.037)

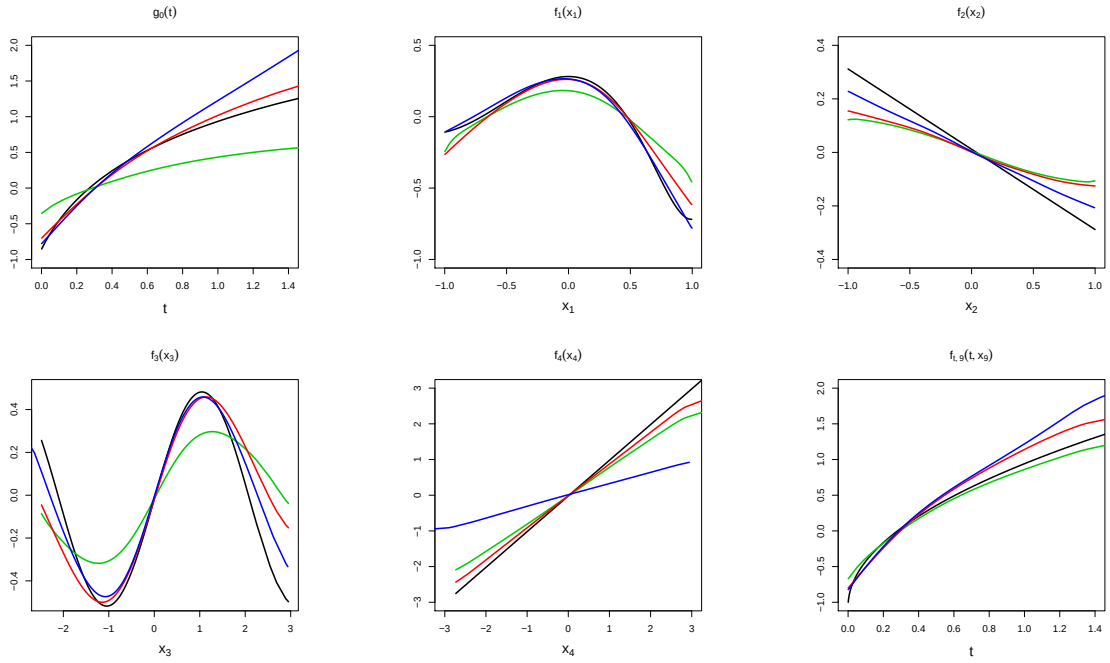

Figure 5: Simulation study. $n = 500$, 40% censoring. 15 covariates (medium noise level). Estimated mean effects along the 200 trials of the continuous covariates included in the theoretical linear predictor. The black line represent the true effects and the colored ones the estimated effects. Blue lines: Two-stage stepwise method. Red lines: Double penalty method. Green lines: Boosting method.

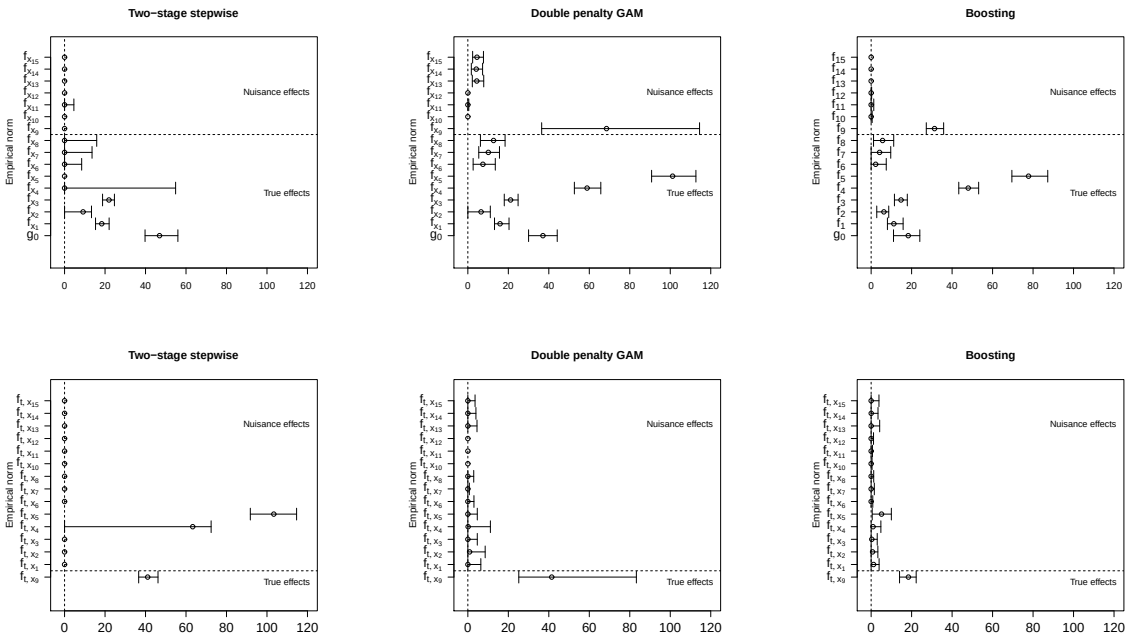

Figure 6: Simulation study. $n = 500$, 40% censoring. 15 covariates (medium noise level). Comparison of model choice performance among methods. The median of the empirical norm along the 200 trials and the first and third quartile are represented for each of the model components included in the three methods. Top panel: Main effects. Bottom panel: Time-varying effects. In each plot, a dashed line separates truly informative components from the nuisance components.

Supplemental scenario 2. $n = 250$, 40% of censored observations. 30 covariates included in each model.

In this scenario, the level of noise is increased by augmenting the number of non-informative covariates. Specifically, the models include 30 covariates, 9 informative and 21 non-informative. $x_1, \ldots, x_{15}$ follow the same specifications than the simulation presented in the main text and supplemental scenario 1.

$x_1, \ldots, x_9$ are informative covariates and were sampled as follows: x_1 to x_4 are continuous while x_5 to x_9 are binary. Specifically $x_1, x_2 \overset{\text{i.i.d.}}{\sim} U(-1, 1)$, $x_3, x_4 \overset{\text{i.i.d.}}{\sim} N(0, 1)$, $x_5, x_9 \overset{\text{i.i.d.}}{\sim} B(1, 0.5)$, and x_6, x_7, x_8 are correlated binary covariates ($\phi(x_i, x_j) = 0.75$; $i, j = 6, 7, 8$). They are the result of the categorization (using the mean as cut-off point) of three variables sampled from a multivariate normal distribution $N_3(\mu, \Sigma)$, where $\mu = (0, 0, 0)'$ and variances in Σ are defined by a parameter $\rho = 0.9$ which governs the level of correlation among covariates.

$x_{10}, \ldots, x_{30}$ are nuisance covariates and were sampled as follows: $x_{10} - x_{12}$ are continuous correlated covariates sampled from a multivariate normal distribution and $x_{13} - x_{15}$ are binary independently draw from $B(1, 0.5)$. $x_{16} - x_{19}$ are correlated binary covariates in the same manner that $x_6 - x_8$, $x_{20} - x_{22}$ are independently sampled from $B(1, 0.5)$, $x_{23} - x_{25}$ are independently draw from $U(-1, 1)$ and $x_{26} - x_{30}$ re derived from a multivariate normal analogously to $x_{10} - x_{12}$.

Table 7: Estimated median MSE and interquartile range (IQR) along 200 trials of the three methods.$n = 250$, 40% censoring. 30 covariates (high noise level)

Theoretical quantity	Two-Stage stepwise	Double Penalty GAM	Boosting
η	3.218(1.315)	2.795(0.891)	2.392(0.924)
g_0	0.337(0.670)	0.148(0.228)	0.244(0.246)
f_1	0.032(0.055)	0.032(0.041)	0.044(0.036)
f_2	0.028(0.024)	0.028(0.020)	0.026(0.020)
f_3	0.041(0.050)	0.033(0.039)	0.116(0.034)
f_4	0.809(0.938)	0.034(0.145)	0.100(0.109)
f_5	2.361(2.492)	0.058(0.254)	0.354(0.463)
f_6	0.011(0.001)	0.059(0.169)	0.023(0.012)
f_7	0.020(0.009)	0.086(0.301)	0.040(0.031)
f_8	0.030(0.004)	0.076(0.205)	0.054(0.047)
$f_{t,9}$	0.136(0.237)	0.076(0.108)	0.061(0.098)

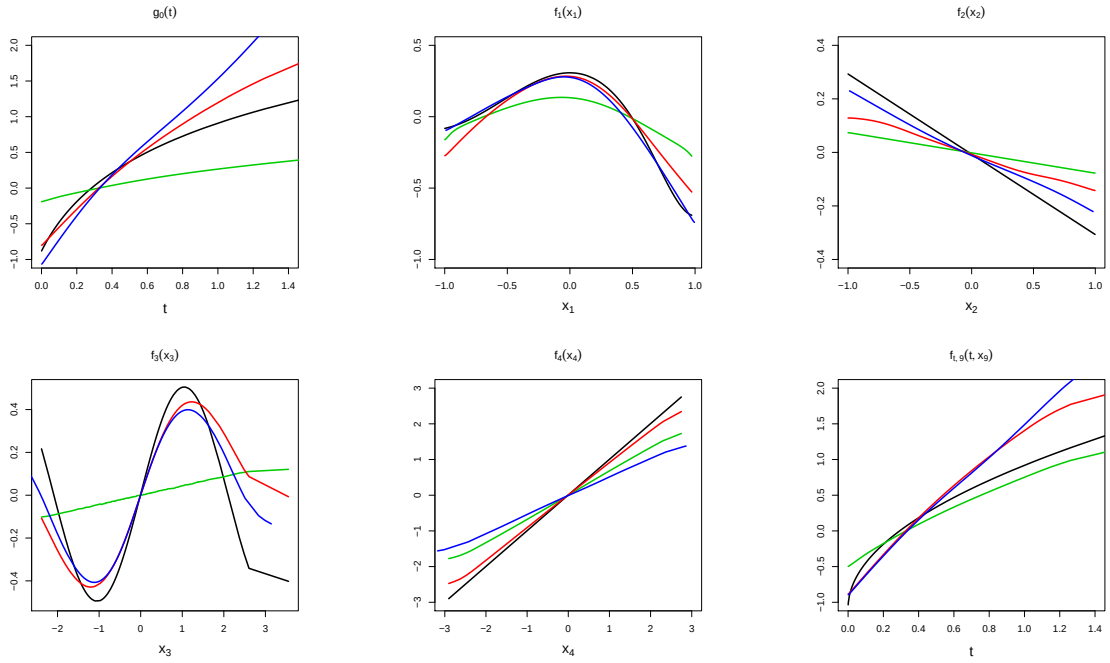

Figure 7: Simulation study. $n = 250$, 40% censoring. 30 covariates (high noise level). Estimated mean effects along the 200 trials of the continuous covariates included in the theoretical linear predictor. The black line represent the true effects and the colored ones the estimated effects. Blue lines: Two-stage stepwise method. Red lines: Double penalty method. Green lines: Boosting method.

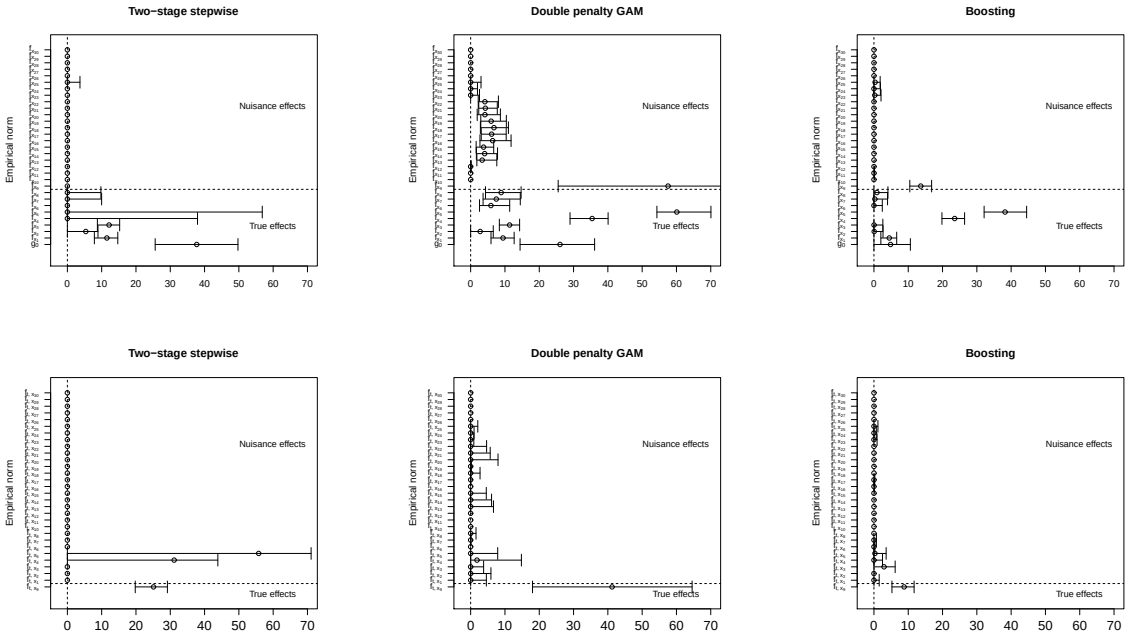

Figure 8: Simulation study. $n = 250$, 40% censoring. 30 covariates (high noise level). Comparison of model choice performance among methods. The median of the empirical norm along the 200 trials and the first and third quartile are represented for each of the model components included in the three methods. Top panel: Main effects. Bottom panel: Time-varying effects. In each plot, a dashed line separates truly informative components from the nuisance components.