LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Susanne Konrath & Thomas Kneib & Ludwig Fahrmeir

Bayesian Regularisation in Structured Additive Regression Models for Survival Data

Technical Report Number 35, 2008
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Bayesian Regularization and Smoothing Priors for Survival Data

Susanne Konrath, Thomas Kneib, Ludwig Fahrmeir

Department of Statistics, Ludwig-Maximilians-University Munich

ABSTRACT:

During recent years, penalized likelihood approaches have attracted a lot of interest both in the area of semiparametric regression and for the regularization of high-dimensional regression models. In this paper, we introduce a Bayesian formulation that allows to combine both aspects into a joint regression model with a focus on hazard regression for survival times. While Bayesian penalized splines form the basis for estimating nonparametric and flexible time-varying effects, regularization of high-dimensional covariate vectors is based on scale mixture of normals priors. This class of priors allows to keep a (conditional) Gaussian prior for regression coefficients on the predictor stage of the model but introduces suitable mixture distributions for the Gaussian variance to achieve regularization. This scale mixture property allows to device general and adaptive Markov chain Monte Carlo simulation algorithms for fitting a variety of hazard regression models. In particular, unifying algorithms based on iteratively weighted least squares proposals can be employed both for regularization and penalized semiparametric function estimation. Since sampling based estimates do no longer have the variable selection property well-known for the Lasso in frequentist analyses, we additionally consider spike and slab priors that introduce a further mixing stage that allows to separate between influential and redundant parameters. We demonstrate the different shrinkage properties with three simulation settings and apply the methods to the PBC Liver dataset.

KEY WORDS:

Bayesian Lasso; hazard regression; Laplace prior; penalized splines; regularization priors; scale mixtures of normals.

1 Introduction

In recent years, penalization approaches have emerged as a general tool that allows to address different problems in applied regression analyses. On the one hand, penalization has been considered for regularizing regression models with a large number of covariates, where penalization introduces shrinkage of estimated coefficients towards zero (e.g. Goeman, 2007, Park and Hastie, 2006 or Efron, Hastie, Johnstone and Tibshirani, 2004). The ultimate goal is to separate between important, influential variables and nuisance covariates that are not associated with the response. On the other hand, smoothness penalties have a long tradition in semiparametric regression, with smoothing splines and penalized polynomial splines as the most prominent examples (see Wood, 2006 or Ruppert, Wand and Carroll, 2003 for overviews). In this case, the penalty represents a roughness measure for unknown functions that avoids overly flexible function estimates. In this paper, we introduce a unifying Bayesian perspective and general Markov chain Monte Carlo (MCMC) simulation algorithms that allow to combine both regularization and smoothing into a general framework. This unifying concept is applied to hazard regression models for continuous time survival analyses based on either the full or the partial likelihood but can also be applied to other types of regression models such as exponential family regression.

From a Bayesian perspective, adding a penalty term to the likelihood corresponds to the assignment of an informative prior distribution to the regression coefficients. More specifically, the penalty term coincides with the negative log-prior, leading to the equivalence of penalized likelihood and posterior mode estimates. For example, the Bayesian analogue of the quadratic ridge penalty is an i.i.d. Gaussian prior, whereas an i.i.d. Laplace prior corresponds to the Lasso (Park and Casella, 2008). The squared difference penalty typically applied in penalized spline smoothing (Eilers and Marx, 1996) relates to a Gaussian random walk assumption for the polynomial spline coefficients (Lang and Brezger, 2004, Brezger and Lang, 2006). For Gaussian prior distributions, efficient proposal densities for exponential family and hazard regression can be derived based on iteratively weighted least squares (IWLS) proposals as introduced by Gamerman (1997) in the context of random effects models (see Brezger and Lang, 2006 for exponential family regression and Hennerfeind, Brezger and Fahrmeir, 2006 for hazard regression). Since the density of the Gaussian prior for the regression coefficients is differentiable, the corresponding full conditionals can be approximated with a Taylor series expansion. The general idea of IWLS proposals is then to obtain a Gaussian proposal by matching the mode and the curvature of the full conditional based on the Taylor expansion. This proposal has two advantages: Firstly, it can be used with multivariate coefficient vectors to take correlations into account in the proposals and, secondly, it automatically adapts to the form of the full conditional thereby avoiding manual tuning of the proposal densities.

Park and Casella (2008) demonstrate how a convenient feature of the Laplace prior allows to construct a Gibbs sampler for Gaussian response models. The Laplace prior can be represented as a scale mixture of normals prior with an exponential mixing hyperprior on the variance. This leads to a hierarchical prior formulation, where Gaussian priors are assigned to the regression coefficients whereas adapted hyperpriors for the variances induce the desired regularization properties. We will employ this representation to extend IWLS proposals in non-Gaussian regression models to spiked regularization priors like the Lasso. Note that the scale mixture of normals class is actually quite large, as demonstrated for example in Griffin and Brown (2005) and other types of regularization priors than the Lasso may be considered in the same framework.

We will employ such extended scale mixtures to address one inherent difficulty with sampling based estimation of regularized regression coefficients: Since estimation is based on samples from the posterior, the posterior mean or median are typically used as point estimates whereas the posterior mode is not available from the samples. As a consequence, estimates obtained with MCMC do no longer have the sharp selection effect that sets some coefficients exactly to zero, a property that led to the popularity of the Lasso penalty. This drawback can be circumvented with Bayesian variable selection schemes that introduce auxiliary binary indicators for non-zero coefficients or non-zero variances (for example Smith and Kohn, 1996, Clyde and George, 2000 or Panagiotelis and Smith, 2008). This leads to discrete-continuous-mixture priors where the discrete mixture component is a point mass in zero. The Bayesian variable selection scheme is particularly attractive in Gaussian regression models or models with a latent Gaussian structure (such as probit models) since efficient marginal samplers can be constructed for the binary indicators in this case.

Instead of a discrete-continuous mixture, Ishwaran and Rao (2003, 2005) introduce Gaussian regression models with a mixture prior of two continuous components (the spike and slab prior) that mimics the Bayesian variable selection idea. One component is a very spiked component that approximates the point mass in zero, whereas the other component is rather flat and noninformative, i. e. corresponds to the slab. This approach has the advantage to ease sampling since no discrete component is involved, but still has the convenient property that small (i. e. practically zero) and larger coefficients can be separated by the mixture indicator in posterior analyses. Note also that in particular in the context of prediction, Hans (2008) provides some evidence that despite its missing variable selection property, the posterior mean may be more attractive than the posterior mode.

In summary, several regularization and smoothing priors can be cast into a hierarchical representation, where the conditional prior for the regression coefficients is Gaussian with suitable (mixture) hyperpriors for the variance. In Gaussian response models, this property facilitates the construction of Gibbs samplers, at least for the regression coefficients. In this paper, we extend both the Lasso prior and the spike and slab prior of Ishwaran and Rao (2005) from Gaussian regression models to Bayesian hazard regression models by adapting the IWLS proposal scheme developed in Hennerfeind, Brezger and Fahrmeir (2006) for geoadditive survival models. In addition to partial likelihood estimation, we consider a full likelihood specification, where the baseline hazard rate is approximated by a penalized spline. The full likelihood approach has the advantage that it facilitates prediction and combines determination of the baseline hazard rate and the regression coefficients into one single estimation scheme that can also be extended to structured additive predictors including nonparametric and time-varying effects.

The rest of the paper is organized as follows: Section 2 introduces different types of hazard regression models and Bayesian regularization priors. In particular, scale mixture representations of the ridge, the Laplace, and the spike and slab prior will be introduced along with regularization priors for penalized spline smoothing. Section 3 discusses posterior inference based on MCMC simulations. The Sections 4 and 5 are devoted to simulations and applications to demonstrate the flexibility and applicability of the proposed methodology. Finally, the concluding discussion section 6 contains a summary and comments on directions of future research.

2 Bayesian Regularization of Hazard Rate Models

This section extends the classical Cox model in two directions: First, the vector β of covariate effects is high-dimensional, possibly including the $p > n$ paradigm arising in microarray-based survival studies. Second, time-varying or nonlinear effects of further covariates may have to be incorporated. Additionally, a smooth nonparametric estimate of the baseline hazard is of interest in many situations. After introducing the models and corresponding likelihoods, we describe regularization in terms of shrinkage priors to deal with the first issue, and through smoothness priors for nonlinear functional effects for the second issue including the baseline hazard as special case.

2.1 Survival models and likelihoods

Let right censored survival data be given in usual form by

$$\mathfrak{D} = \left\{\left(t_i, d_i, x_i'\right), i = 1, ..., n\right\},$$

where $t_i = \min(T_i, C_i)$ is the observed lifetime, i. e. either the time until death or the time until the end of the observation (in the right censoring case the event did not happen before the individual is censored). With $d_i = I(T_i \leq C_i)$ we denote the censoring indicator, and $x_i' = (x_{i1}, ..., x_{ip})$ is a vector of time independent covariates. We consider noninformative censoring, with independent lifetimes T_i and censoring times C_i. Additionally we assume that continuous covariates are standardized in advance, thus avoiding adjustment of shrinkage priors for different covariate scales.

In Cox's proportional hazard model (Cox, 1972), the hazard rate for individual i is

$$\lambda_i\left(t\right) = \lambda_0\left(t\right)\exp\left(x_i'\beta\right). \tag{2.1}$$

The baseline hazard $\lambda_0(t)$ is left unspecified and, for small p, the parameters $\beta = (\beta_1, ..., \beta_p)'$ are usually estimated via maximization of the partial likelihood. The Breslow estimate may be computed in a second step yielding a step function for the cumulative baseline hazard (e. g. Lin, 2007).

For full Bayesian inference, we include the log-baseline hazard $g_0\left(t\right) = \log\lambda_0\left(t\right)$ into the predictor, resulting in

$$\lambda_i\left(t\right) = \exp\left(g_0\left(t\right) + x_i'\beta\right) = \exp\left(\eta_i\left(t\right)\right). \tag{2.2}$$

Specifying, for example, $g_0(t)$ through a regression spline with a smoothness prior for the basis coefficients (see subsection 2.3), joint inference for covariate effects and the baseline hazard based on the full likelihood becomes feasible. Obtaining a full probabilistic framework is a useful feature if modelling of individual hazards and predictions are of interest.

The semiparametric predictor $\eta_i(t)$ in (2.2) can be further extended to

$$\lambda_i(t) = \exp\left(g_0(t) + x_i'\beta + u_i'\zeta + \sum_{j=1}^{p} g_j(t) v_{ij} + \sum_{j=1}^{q} f_j(z_{ij})\right) = \exp(\eta_i(t)) \qquad (2.3)$$

where u_i is a vector consisting of a small number of additional 'usual' covariates, such as sex or age of a patient, with linear effects ζ that should not be regularized. The functions $g_j(t)$ are time-varying effects of covariates v_j, and $f_j(z_j)$ are the nonlinear effects of continuous covariates z_j. Further components, such as nonlinear interactions between two continuous covariates, spatial effects and group-specific effects may also be included. Cox-type models with such general forms of structured additive predictors have been suggested in Hennerfeind et al. (2006), however without considering the problem of high-dimensional β. By incorporating time-varying covariate effects, the proportional hazards assumption of the Cox model is relaxed. In addition possibly nonlinear effects of continuous covariates, for example age of a patient, can be explored. In this paper we show how the $p > n$ paradigm and semiparametric inference for functional effects can be treated within a unifying framework because due to the regularization priors there is no need to distinguish between the cases $p \leq n$ and $p > n$ conceptionally, if all regression coefficients are equipped with more or less strong regularization prior.

It turns out that the vector $\eta = (\eta_1, ..., \eta_n)$ of predictors $\eta_i = \eta_i(t_i)$ in (2.2) or (2.3), evaluated at observed lifetimes $t_i, i = 1, ..., n,$ can always be represented, after reindexing, as a high-dimensional predictor

$$\eta = X\beta + U\zeta + Z_0\gamma_0 + ... + Z_m\gamma_m .$$

The design matrices X and U have rows x_i' and u_i', the design matrices $Z_0, ..., Z_m$ are constructed from basis functions representing the functions $g_0, ..., g_q, f_1, ..., f_r,$ and $\gamma_0, ..., \gamma_m$ are corresponding vectors of basis coefficients, see subsection 2.3. The resulting inverse problem of estimating the parameters will be regularized through informative shrinkage priors for the components of β, a flat or weakly informative Gaussian prior for ζ, and Gaussian smoothness priors for $\gamma_0, ..., \gamma_m$ in the following subsections.

Assuming noninformative right-censoring, the full likelihood is given by

$$L(\lambda) = \prod_{i=1}^{n} \lambda_i(t_i)^{d_i} \exp\left(-\int_0^{t_i} \lambda_i(u) du\right), \qquad (2.4)$$

inserting $\lambda_i(t_i)$ and the expressions for the predictors. For the predictor in (2.2), we obtain

$$L(\beta,\lambda_0) = \exp\sum_{i=1}^{n}\left[d_i\left(\log\lambda_0(t_i) + x_i'\beta\right) - \exp(x_i'\beta)\int_0^{t_i}\lambda_0(u)\,du\right] = \exp\left(l(\beta,\lambda_0)\right). \qquad (2.5)$$

as a special case. Apart from simple parametric forms of the baseline hazard rate, for example a Weibull model or a piecewise constant function, the integral has to be evaluated numerically, using, e.g., the trapezoidal rule as in Hennerfeind et al. (2006).

If primary interest is on β for model (2.2), without specification of the baseline hazard, estimation is usually carried out by maximization of the partial likelihood

$$pL(\beta) = \prod_{i=1}^{n}\left\{\frac{\exp(x_i'\beta)}{\sum_{k=1}^{n} I(t_k \geq t_i)\exp(x_k'\beta)}\right\}^{d_i} = \exp\left(pl(\beta)\right) \qquad (2.6)$$

where pl denotes the logarithm of the partial likelihood. The indicator function in the denominator is used to describe if individual k is still under risk at time point t_i^-. The partial likelihood only depends on the order of the failure times not on the exact values of failure times. Modifications are required if tied failure times are present. For simplicity we assume no ties and refer e. g. to Therneau and Grambsch (2000) for corrections to handle with ties.

For Bayesian inference it seems questionable if the partial likelihood can be used instead of the genuine full likelihood (2.4) or (2.5) for posterior analysis. Sinha et al. (2003) provide a rigorous justification for model (2.2), when the (cumulative) baseline hazard is specified through a gamma process prior. We will instead specify $\lambda_0(t)$ through a log-normal process prior. Because gamma and log-normal process priors are closely related from a practical point of view, we argue heuristically that Bayesian inference can again be based on the partial likelihood. Section 4 provides empirical evidence for this conjecture.

2.2 Shrinkage Priors

To deal with the problem of variable selection and regularzation we consider and compare several shrinkage priors. Some of these priors correspond to well-known frequentist shrinkage penalties, such as the Lasso or Ridge penalty. A desirable feature of shrinkage priors used for variable selection is to shrink small effects close to zero, but to shrink significant effects only moderately to prevent them from large bias, see the discussion in Griffin and Brown (2005, 2007) or Zou (2006). All priors considered in the following sections can be represented as scale mixtures of normal priors, which is very useful for MCMC inference.

2.2.1 Ridge prior

A well known penalty to deal with multicollinearity or the problem of $p > n$ in classical regression is the Ridge penalty. The Bayesian version of the Ridge penalty is given by the assumption of i.i.d. Gaussian priors for the regression coefficients

$$\beta_j \mid \lambda \sim_{\text{iid}} N\left(0, 1/2\lambda\right), \quad j = 1, \ldots, p,$$

that leads to the prior density

$$\pi\left(\beta \mid \lambda\right) = \prod_{j=1}^{p} \pi\left(\beta_j \mid \lambda\right) = \left(\sqrt{\frac{\lambda}{\pi}}\right)^p \exp\left\{-\lambda \sum_{j=1}^{p} \beta_j^2\right\},$$

with the scale mixture representation

$$\beta_j \mid \tau_j^2 \sim N\left(0; \tau_j^2\right), \quad \tau_j^2 \mid \lambda \sim \delta_{1/2\lambda}\left(\tau_j^2\right).$$

The symbol $\delta_a(t)$ denotes the Kronecker function which is 1 if $t = a$ and 0 if $t \neq a$. For given λ posterior mode estimation corresponds to penalized likelihood estimation. Due to conjugacy to the Gaussian family, an additional gamma prior is used for the shrinkage parameter

$$\lambda \sim \text{Gamma}\left(a_\lambda, b_\lambda\right); \quad a_\lambda, b_\lambda > 0$$

that supports a Gibbs update for this parameter. The marginalisation over λ results in an inverse gamma distribution for the variance parameters

$$\tau^2 \mid a_\lambda, b_\lambda \sim \text{InvGamma}\left(a_\lambda, \frac{1}{2} b_\lambda\right)$$

and further marginalization in a scaled t distribution for the marginal distribution of the regression coefficients given the hyperparameters a_λ, b_λ

$$\pi\left(\beta_j \mid a_\tau, b_\tau\right) = \int_0^\infty N\left(\beta_j \mid 0, \tau_j^2\right) \text{InvGamma}\left(\tau_j^2 \mid a_\lambda, \frac{b_\lambda}{2}\right) d\tau_j^2$$

$$= t\left(\beta_j \mid df = 2a_\tau, \text{scale} = \sqrt{b_\tau / 2a_\tau}\right).$$

The additional prior assumption about the shrinkage parameter leads to a more flexible modelling of our prior knowledge and a refinement of the prior tuning in order to shrink the parameters via the two hyperparameters in $\pi\left(\beta_j \mid a_\tau, b_\tau\right)$ compared to the normal prior $\pi\left(\beta_j \mid \lambda\right)$. Besides, we get another method to determine the shrinkage parameter, via the mean or the median of the posterior sample of the marginal shrinkage parameter. Compared to the crossvalidation methods in “classical” penalized regression, the Bayesian approach provides a very simple access to an estimate $\hat{\lambda}$, especially compared to the burden crossvalidation can cause for complex models.

2.2.2 Lasso prior

Just as well known as Ridge regression is the Lasso regression (Tibshirani, 1996) if simultaneously variable selection and estimation should be done. The Bayesian version of the Lasso can be formulated with i.i.d. Laplace priors

$$\beta_j \mid \lambda \sim_{\text{iid}} \text{Laplace}\left(0, \lambda\right), \quad j = 1, \ldots, p, \tag{2.7}$$

with common density

$$p\left(\beta \mid \lambda\right) \propto \exp\left(-\lambda \sum_{j=1}^{p} \left|\beta_j\right|\right)$$

compare e. g. Park and Casella (2008). Figure.2.1 shows the Laplace prior in the univariate case. This is the well-known Lasso penalty, and as in Ridge regression - for given λ - posterior mode estimation corresponds to penalized likelihood estimation. For full Bayesian inference, it is again convenient to express the Laplace density as a scale mixture of normals introducing a further stage in the hierarchical model formulation:

$$\beta_j \mid \tau_j^2 \sim N\left(0; \tau_j^2\right), \quad \tau_j^2 \mid \lambda^2 \sim_{\text{iid}} \text{Exp}\left(\frac{\lambda^2}{2}\right). \tag{2.8}$$

To obtain a data driven penalty we additionally use a gamma prior for the squared shrinkage parameter λ^2, i. e.

$$\lambda^2 \sim \text{Gamma}\left(a_\lambda, b_\lambda\right); \quad a_\lambda, b_\lambda > 0 .$$

This hierarchy leads to the following density of the marginal distributions for the variance parameter

$$\pi\left(\tau_j^2 \mid a_\lambda, b_\lambda\right) = \int \text{Exp}\left(\tau_j^2 \mid \frac{\lambda^2}{2}\right) \text{Gamma}\left(\lambda^2 \mid a_\lambda, b_\lambda\right) d\lambda^2 = \frac{a_\lambda}{2b_\lambda}\left[\frac{\tau_j^2}{2b_\lambda} + 1\right]^{-\left(a_\lambda + 1\right)} .$$

We denote this distribution as $\text{ExpGamma}\left(\tau_j^2 \mid a_\lambda, b_\lambda\right)$. The marginal density of the regression coefficients (Figure.2.1) can be expressed as

$$\begin{aligned}\pi\left(\beta_j \mid a_\lambda, b_\lambda\right) &= \int N\left(\beta_j \mid 0, \tau_j^2\right) \text{ExpGamma}\left(\tau_j^2 \mid a_\lambda, b_\lambda\right) d\tau_j^2 \\ &= \frac{a_\lambda}{\sqrt{\pi}} \frac{2^{a_\lambda}}{2b_\lambda} \Gamma\left(a_\lambda + 1/2\right) \exp\left(\frac{1}{4} \frac{\beta_j^2}{\left[2b_\lambda\right]^2}\right) D_{-2\left(a_\lambda + 1/2\right)}\left(\frac{\left|\beta_j\right|}{2b_\lambda}\right)\end{aligned}$$

with the parabolic cylinder Function $D_{-2a_\lambda - 1}$, see Griffin and Brown (2005, 2007). As mentioned above in the Ridge penalty section, we also get a further method to estimate the shrinkage parameter and a more flexible prior for the regression coefficients.

2.2.3 NMIG prior

As a further mixture prior we consider a normal mixture of inverse gamma distributions, shortly named as NMIG prior. This prior has been suggested for regularizing high-dimensional linear Gaussian regression models by Ishwaran and Rao (2003, 2005). The conditional Gaussian distribution for the regression coefficients is Gaussian as in the Lasso and Ridge case, i. e.

$$\beta_j \mid I_j, \psi_j^2 \sim N\left(0; \tau_j^2 := I_j \psi_j^2\right),$$

but the variance parameters τ_j^2 of this distribution are in contrast assigned a mixture distribution modelled through the product of the two components

$$\begin{aligned} I_j \mid \nu_0, \nu_1, \omega &\sim (1-\omega)\delta_{\nu_o}(\cdot) + \omega\delta_{\nu_1}(\cdot) \\ \psi_j^2 \mid a_\psi, b_\psi &\sim \text{InvGamma}\left(a_\psi, b_\psi\right). \end{aligned} \tag{2.9}$$

The first component in (2.9) is an indicator variable with point mass at the values $\nu_0 > 0$ and $\nu_1 > 0$ denoted by the corresponding Kronecker symbols. Therein the parameter ν_0 should have a positive value close to zero and the value of ν_1 is 1. The parameter ω controls how likely the binary variable I_j equals ν_1 or ν_0, and therefore it takes on the role of a complexity parameter that controls the size of the models. The assumptions in (2.9) are leading to a continuous bimodal distribution for the variances parameter $\tau_j^2 := I_j \psi_j^2$, given $\nu_0, \nu_1, \omega, a_\psi, b_\psi$ as a mixture of scaled inverse Gamma distributions

$$\pi\left(\tau_j^2 \mid \nu_0, \nu_1, \omega, a_\psi, b_\psi\right) = (1-\omega) \cdot \text{InvGamma}\left(\tau_j^2 \mid a_\psi, \nu_0 b_\psi\right) + \omega \cdot \text{InvGamma}\left(\tau_j^2 \mid a_\psi, \nu_1 b_\psi\right).$$

We assume a uniform prior for the parameter ω to express an indifferent prior knowledge about the model complexity

$$\omega \sim \text{Uniform}(0,1).$$

To transport more information into the model it is possible to use a beta prior $\text{Beta}(a_\omega, b_\omega)$ for ω, which reduces to the uniform prior in the special case $a_\omega = b_\omega = 1$. With an appropriate choice of the hyperparameters $a_\omega, b_\omega > 0$ we are able to favour more or less sparse models. Apart from the special choice of the prior for ω the use of a continuous prior for ω has several advantages than using a degenerate prior as for example in George and McCulloch (1993) or Geweke (1996). First, the update of the variance parameter components can easily be done via Gibbs sampling and no complicated updates are necessary, compare George and McCulloch (1997) for variable selection priors with a point mass at zero for linear models. Furthermore, it is possible to select important model variables via the posterior mean of the corresponding indicators I_j and to simultaneously estimate their values like in the Lasso case. Finally, the uniform prior allows for a greater amount of adaptiveness in estimating the model size. In addition, the estimates for relevant covariates should be less biased than in the case of unimodal priors for the regression coefficients because the bimodality supports less penalization of large coefficients.

The marginal density for the variance parameters, after integrating out ω is the mixture

$$\pi\left(\tau_j^2 \mid \nu_0,\nu_1,a_\psi,b_\psi\right)=0.5\cdot\text{InvGamma}\left(\tau_j^2 \mid a_\psi,\nu_0 b_\psi\right)+0.5\cdot\text{InvGamma}\left(\tau_j^2 \mid a_\psi,\nu_1 b_\psi\right),$$

which corresponds to the conditional density $\pi\left(\tau_j^2 \mid \nu_0,\nu_1,\omega,a_\psi,b_\psi\right)$ for the choice $\omega=0.5$. The prior locations of the two modes are independent of ω and fixed at

$$\text{mod}\,e_{\nu_0}=\frac{\nu_0 b_\psi}{a_\psi+1},\quad \text{mod}\,e_{\nu_1}=\frac{\nu_1 b_\psi}{a_\psi+1}.$$

The marginal density for the regression coefficients is given by

$$\pi\left(\beta_j \mid \nu_0,\nu_1,a_\psi,b_\psi\right)=\int_0^\infty N\left(\beta_j \mid 0,\tau_j^2\right)\pi\left(\tau_j^2 \mid \nu_0,\nu_1,a_\psi,b_\psi\right)d\tau_j^2$$

$$=\frac{1}{2}\frac{\Gamma\left(\frac{2a_\psi+1}{2}\right)}{\Gamma\left(\frac{2a_\psi}{2}\right)\sqrt{2a_\psi\frac{\nu_0 b_\psi}{a_\psi}\pi}}\left(1+\frac{\beta_j^2}{2a_\psi\frac{\nu_0 b_\psi}{a_\psi}}\right)^{-\frac{2a_\psi+1}{2}}+\frac{1}{2}\frac{\Gamma\left(\frac{2a_\psi+1}{2}\right)}{\Gamma\left(\frac{2a_\psi}{2}\right)\sqrt{2a_\psi\frac{\nu_1 b_\psi}{a_\psi}\pi}}\left(1+\frac{\beta_j^2}{2a_\psi\frac{\nu_1 b_\psi}{a_\psi}}\right)^{-\frac{2a_\psi+1}{2}}$$

so that the marginal distribution for the components of β is a mixture of scaled t-distributions

$$\beta_j \mid \nu_0,a_\psi,b_\psi \sim \frac{1}{2}t\left(\beta_j \mid df=2a_\psi,\text{scale}=\sqrt{\frac{\nu_0 b_\psi}{a_\psi}}\right)+\frac{1}{2}t\left(\beta_j \mid df=2a_\psi,\text{scale}=\sqrt{\frac{\nu_1 b_\psi}{a_\psi}}\right).$$

2.2.4 Adaptive priors

To achieve more flexibility, we can equip the models above with separate complexity parameters. The resulting models are additionally named with “adaptive”. For example, the adaptive version of the Lasso prior is given through $\tau_j^2 \mid \lambda_j^2 \sim_{\text{iid}} \text{Exp}\left(\lambda_j^2/2\right)$, $\lambda_j^2 \sim_{\text{iid}} \text{Gamma}\left(a_\lambda,b_\lambda\right)$ and the adaptive NMIG prior with $\omega_i \sim \text{Uniform}\left(0,1\right)$ or $\omega_j \sim \text{Beta}\left(a_\omega,b_\omega\right)$. It is straightforward to use individual hyperparameters, i. e. $a_{\lambda_j},b_{\lambda_j}$ in the Lasso case and $a_{\omega_j},b_{\omega_j}$ in the NMIG case. This can be done if the covariates are not standardized to take account for different scales. However, one should keep in mind that the number of parameters to estimate is increasing in the adaptive versions, which can cause problems in situations with low sample sizes.

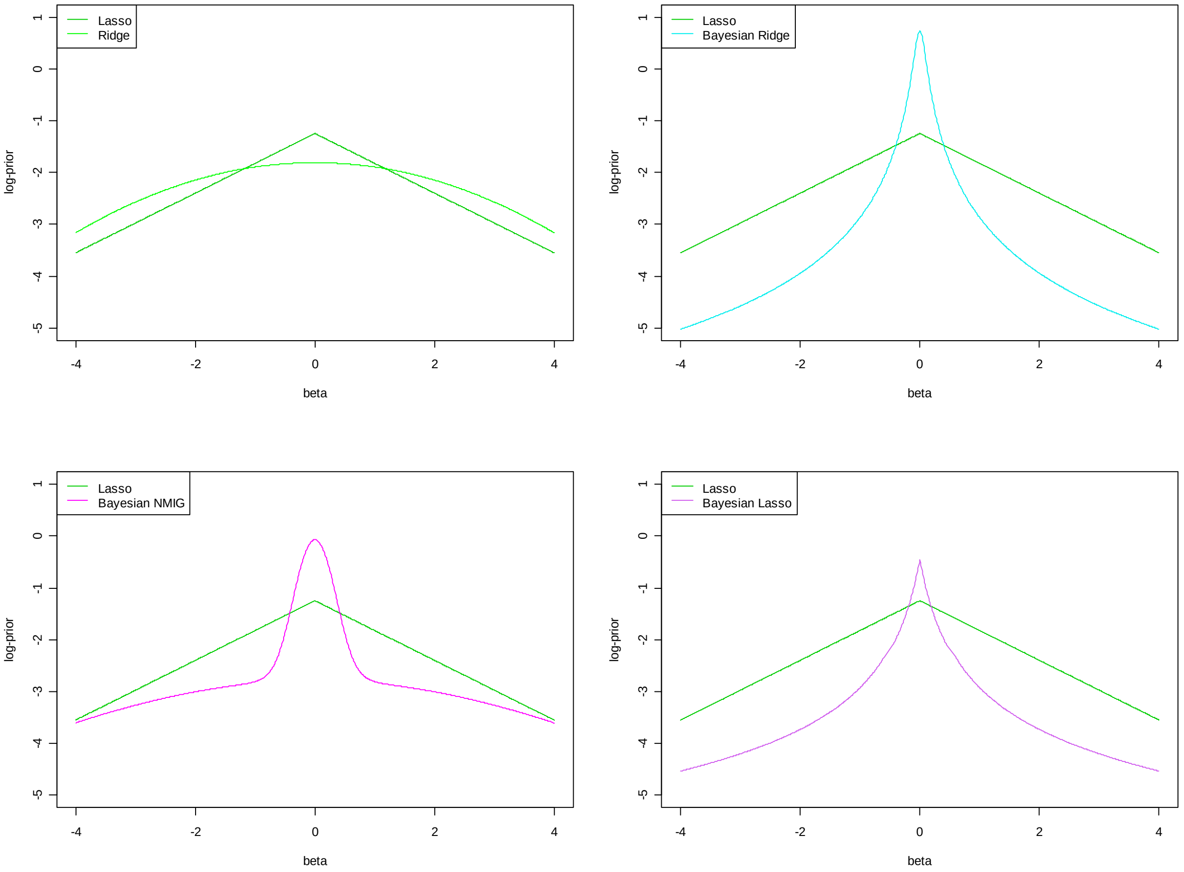

Figure 2.1: Log-priors in comparison to the Lasso prior (green line).

2.2.5 Variable selection

Since the Bayesian regularization priors do not share the strong variable selection property of the frequentist Lasso, hard shrinkage rules are considered to accomplish variable selection. A first rule is based on the 95% credible intervals (CI95), obtained from the corresponding sample quantiles of the regression coefficient's MCMC samples. A second interval criterion is constructed using the sample standard deviation, so that only regression coefficients with zero outside the one standard deviation interval around the posterior mean are included in the final model. On the other hand, the Bayesian NMIG provides a natural criterion to select covariates if the samples of the indicator variables are utilized. Covariates with considerable influence should be assigned to the mixing distribution component corresponding to the indicator with values $\nu_1 = 1$. The more the posterior mean of an indicator variable increases (i. e. the percentage of the values $\nu_1 = 1$ in the sample), the larger is the evidence that the corresponding covariate has non negligible effect. In our simulations and application we use the intuitive cut value of 0.5 as a selection criterion, i. e. covariates whose corresponding indicator posterior mean exceeds 0.5 are included in the final model, see Sections 4 and 5.

2.3 Smoothness priors

Smooth modelling and estimation of nonlinear and time-varying effects including the (log-) baseline hazard, is based on Bayesian P-splines. If we use the partial likelihood for Bayesian inference in model (2.1), no additional prior assumption for the baseline hazard is needed. A justification for using the partial likelihood instead of a genuine likelihood in a Bayesian setting is given in Sinha et al. (2003) for time-constant effects.

A flexible approach to model and to estimate the baseline hazard is to use an approximation with spline basis functions of degree ℓ. To do so, one chooses a sequence of knots $\xi_1 < \ldots < \xi_s$ from $\left(t_{\min}, t_{\max}\right)$ with additional boundary knots $0 = \xi_0 < \xi_1$ and $\xi_s < \xi_{s+1} = \infty$. Denote with $B_1\left(\cdot\right), \ldots, B_M\left(\cdot\right)$, $M = s + \ell - 1$ an appropriate basis of the spline space $S^{\ell}\left(\xi_1, \ldots, \xi_s\right)$, then the baseline hazard is approximated through

$$g_0\left(t\right) = \log\lambda_0\left(t\right) = B_1\left(t\right)\gamma_{01} + \ldots + B_M\left(t\right)\gamma_{0M}.$$

We can express the predictors as

$$\eta_i\left(t_i\right) = \exp\left(\log\lambda_0\left(t_i\right) + x_i'\beta\right) = \exp\left(z_{i0}'\gamma_0 + x_i'\beta\right)$$

with $z_{i0}' = \left(B_1(t_i), \ldots, B_M(t_i)\right)$, and the predictor $\eta = \left(\eta_1, \ldots, \eta_n\right)'$ in the form

$$\eta = Z_0\gamma_0 + X\beta.$$

To guarantee smoothness for the unknown log baseline hazard $g_0(t)$, we assume Bayesian P-spline priors as in Lang and Brezger (2004). This implies the use of B-spline basis functions of degree ℓ to model the log-baseline and first or second order random walk smoothness priors for the parameter vector γ_0,

$$\gamma_{0,m} = \gamma_{0,m-1} + u_{0,m} \quad \text{or} \quad \gamma_{0,m} = 2\gamma_{0,m-1} + \gamma_{0,m-2} + u_{0,m}$$

with i.i.d. Gaussian errors $u_{0,m} \sim N(0,\delta_0^2)$ and diffuse priors for the initial values $\pi(\gamma_{0,1}) \propto \text{const}$ or $\pi(\gamma_{0,1}) = \pi(\gamma_{0,2}) \propto \text{const}$. The first order random walk prior controls abrupt jumps in the differences $\gamma_{0,m} - \gamma_{0,m-1}$, while the second order random walk prior penalizes deviations from a linear trend. The variance parameter δ_0^2 controls the amount of the penalization and acts as a smoothness parameter. The smaller the variance parameter, the stronger is the penalization. The joint prior for the parameter γ_0 as the product of the conditional densities is

$$\pi\left(\gamma_0 \mid \delta_0^2\right) \propto \left(\frac{1}{\delta_0^2}\right)^{\frac{k}{2}} \exp\left(-\frac{1}{2\delta_0^2}\gamma_0' K \gamma_0\right)$$

with penalty matrix K of the form $K = D'D$, where D is a first or second order difference matrix and $k = \text{rank}\left(K\right)$. A standard option for the variance parameter is a diffuse inverse gamma prior $\delta_0^2 \sim \text{InvGamma}(a_0, b_0)$ with density

$$\pi\left(\delta_0^2 \mid a_0, b_0\right) \propto \frac{1}{\left(\delta_0^2\right)^{a_0+1}} \exp\left(-\frac{b_0}{\delta_0^2}\right),$$

and small $a_0 > 0,\ b_0 > 0$. It is also possible to choose an improper Gamma-type prior in the case when $a_0 \le 0,\ b_0 \le 0$ or especially $a_0 \le 0,\ b_0 = 0$, for example $a_0 = -0.5$, $b_0 = -1$, corresponding to a flat prior $p(\delta_0) \propto \text{const}$ for the standard deviation $\delta_0 = \sqrt{\delta_0^2}$.

In the extended model (2.3), unknown time-varying effects $g_j(t)$ and nonparametric functions $f_j(z_j)$ of continuous covariates are modelled through Bayesian P-splines as well. Let $g_j = \left(g_j(t_1)v_{ij}, \ldots, g_j(t_n)v_{ij}\right)'$ denote the vector of time-varying effects $g_j(t_i)v_{ni}$ and $f_j = \left(f_j(z_{ij}), \ldots, f_j(z_{ni})\right)'$ the vector of evaluations of $f_j(z_j)$. Then g_j and f_j can be expressed as the product of an appropriately defined design matrix Z_j and a (possibly) high-dimensional vector γ_j of parameters, for example $g_j = Z_j\gamma_j$ or $f_j = Z_j\gamma_0$ .After reindexing, we can represent the predictor vector η in generic notation as

$$\eta = X\beta + U\zeta + Z_0\gamma_0 + \ldots + Z_m\gamma_m,$$

see Hennerfeind et al. (2006) for more details and extensions to spatial and group-specific effects. The general form of smoothness priors for γ_j is

$$p(\gamma_j \mid \delta_j^2) \propto \delta_j^{-k_j} \exp(-\frac{1}{2\delta_j^2}\gamma_j' K_j \gamma_j),$$

with precision or penalty matrix K_j and $k_j = \text{rank}(K_j),\ j = 0,1,...,m$.

3 Posterior inference

We first describe MCMC inference for the basic model (2.2) with predictor $\eta_i(t) = g_0(t) + x_i'\beta$, where shrinkage estimates of β, possibly together with a smooth estimate of the log-baseline rate $g_0(t)$, are of interest. Joint shrinkage and smoothing in the extended model (2.3) is outlined subsequently.

3.1 MCMC with shrinkage priors

For inference based on the full likelihood, the posterior has the general form

$$\pi\left(\beta, \gamma_0, \delta_0^2, \tau^2, \phi \mid \mathfrak{D}\right) \propto \exp\left(l\left(\beta, \gamma_0\right)\right) \pi\left(\gamma_0 \mid \delta_0^2\right) \pi\left(\delta_0^2\right) \prod_{j=1}^{p} p\left(\beta_j \mid \tau_j^2\right) \pi\left(\tau^2, \phi\right),$$

where $\tau^2 = (\tau_1^2,...,\tau_p^2)$ and $\pi(\tau^2, \phi)$ is a generic notation for priors defined for τ^2 and other parameters or random variables ϕ through the hierarchical formulation of shrinkage priors in Section 2.2. Thus, for all shrinkage priors, the full conditional of the regression parameters β is

$$\pi\left(\beta \mid \cdot\right) \propto \exp\left\{ l\left(\beta, \gamma_0\right) - \frac{1}{2}\beta' D_\tau^{-1} \beta \right\}$$

where $D_\tau = \text{diag}\left(\tau_1^2,...,\tau_p^2\right)$ denotes the matrix of the variance parameters. This distribution has no closed form to draw a new/proposed state for the Markov chain. We use an MH-algorithm with so-called IWLS-proposals to update the regression coefficients. To do so, the log-likelihood is approximated by a second order Taylor expansion at the current state of the parameter vector $\beta^{(c)}$, i. e.

$$\begin{aligned} l\left(\beta, \gamma_0\right) &\approx l\left(\beta^{(c)}, \gamma_0\right) + \left(\beta - \beta^{(c)}\right)' s_\beta\left(\beta^{(c)}, \gamma_0\right) + \frac{1}{2}\left(\beta - \beta^{(c)}\right)' H_\beta\left(\beta^{(c)}, \gamma_0\right)\left(\beta - \beta^{(c)}\right) \\ &\propto -\frac{1}{2}\beta'\left[-H_\beta\left(\beta^{(c)}, \gamma_0\right)\right]\beta + \beta'\left[s_\beta\left(\beta^{(c)}, \gamma_0\right) - H_\beta\left(\beta^{(c)}, \gamma_0\right)\beta^{(c)}\right] \end{aligned}$$

with Hessian matrix H_β and score vector s_β. The likelihood is then approximated by a multivariate Gaussian distribution with precision matrix and mean

$$\Sigma_\beta = -H_\beta\left(\beta^{(c)}, \gamma_0\right)$$

$$\mu_\beta = \Sigma_\beta^{-1}\left[s_\beta\left(\beta^{(c)}, \gamma_0\right) - H_\beta\left(\beta^{(c)}, \gamma_0\right)\beta^{(c)}\right]$$

The score function is

$$s_\beta(\beta,\gamma_0) = \frac{\partial l(\beta,\gamma_0)}{\partial\beta} = X'\left[d - \Lambda(t\,|\,\beta,\gamma_0)\right]$$

with $\Lambda(t\,|\,\beta,\gamma_0) = \left(\Lambda_0(t_1;\gamma_0)\exp(x_1'\beta), ..., \Lambda_0(t_n;\gamma_0)\exp(x_n'\beta)\right)'$ and $d = (d_1, ..., d_n)'$. The Hessian matrix is

$$H_\beta(t\,|\,\beta,\gamma_0) = \frac{\partial s(t\,|\,\beta,\gamma_0)}{\partial\beta} = -X'\mathrm{diag}\left(\Lambda(t\,|\,\beta,\gamma_0)\right)X\,.$$

The approximation of the full conditional $\pi(\beta\,|\,\cdot)$ through a multivariate Gaussian distribution $\hat{\pi}\left(\cdot\,|\,\beta^{(c)},\cdot\right)$ with precision and mean

$$\Sigma_\beta = X'\mathrm{diag}\left(\Lambda\left(t\,|\,\beta^{(c)},\gamma_0\right)\right)X + D_\tau^{-1}$$

$$\mu_\beta = \Sigma_\beta^{-1}\left[X'd - X'\Lambda\left(t\,|\,\beta^{(c)},\gamma_0\right) + X'\mathrm{diag}\left(\Lambda\left(t\,|\,\beta^{(c)},\gamma_0\right)\right)X\beta^{(c)}\right]$$

is used as the proposal density to draw a new proposal $\beta^{(p)}$ of the Markov chain which is accepted with the probability

$$\alpha\left(\beta^{(p)},\beta^{(c)}\right) = \min\left\{1, \frac{\pi\left(\beta^{(p)}\,|\,\cdot\right)\hat{\pi}\left(\beta^{(c)}\,|\,\beta^{(p)},\cdot\right)}{\pi\left(\beta^{(c)}\,|\,\cdot\right)\hat{\pi}\left(\beta^{(p)}\,|\,\beta^{(c)},\cdot\right)}\right\}.$$

MH steps for updating the baseline coefficients γ_0 can conceptually be carried out in the same way as for β, replacing the (conditional) precision matrix D_τ^{-1} by K/τ_0^2, the precision matrix of the (conditional) Gaussian prior. However, computational efforts increase because the cumulative baseline hazard is the integral of the baseline hazard involved, complicating derivation and computation of the score function s_{γ_0} and the Hessian matrix H_{γ_0}. As an alternative, we may use MH steps with conditional prior proposals as in Hennerfeind et al. (2006).

If we rely on posterior inference based on the partial likelihood, we replace the log-likelihood $l(\beta, \gamma_0)$ through the partial (log-)likelihood $pl(\beta)$ as well as the score function and Hessian matrix through corresponding derivatives. The full conditional for β is then given by

$$\pi(\beta \mid \cdot) \propto \exp\left\{ pl(\beta) - \frac{1}{2}\beta' D_\tau^{-1}\beta \right\},$$

and the (partial) score function and Hessian matrix are

$$ps_\beta(\beta) = \frac{\partial pl(\beta)}{\partial \beta} = \left(\sum_{i=1}^{n} \left(d_i x_{ij} + d_i \frac{\sum_{k=1}^{n} I(t_k \geq t_i)\exp(x_k'\beta)x_{kj}}{\sum_{k=1}^{n} I(t_k \geq t_i)\exp(x_k'\beta)} \right) \right)_{1 \leq j \leq p},$$

$$pH_\beta(\beta) = \left(-\sum_{i=1}^{n} d_i \frac{\sum_{k=1}^{n} I(t_k \geq t_i)\exp(x_k'\beta)x_{kj}x_{km}}{\sum_{k=1}^{n} I(t_k \geq t_i)\exp(x_k'\beta)} \right.$$

$$\left. + \sum_{i=1}^{n} d_i \frac{\sum_{k=1}^{n} I(t_k \geq t_i)\exp(x_k'\beta)x_{kj} \cdot \sum_{k=1}^{n} I(t_k \geq t_i)\exp(x_k'\beta)x_{km}}{\left[\sum_{k=1}^{n} I(t_k \geq t_i)\exp(x_k'\beta)\right]^2} \right)_{1 \leq j,m \leq p}.$$

3.1.1 Bayesian Ridge

The full conditionals for the variance parameters simplify to

$$\tau_j^2 \mid \cdot = \frac{1}{2\lambda} \quad, j = 1,..,p.$$

and the full conditional for the shrinkage parameter is a gamma density,

$$\lambda \mid \cdot \sim \text{Gamma}\left(\frac{p}{2} + a_\lambda, \sum_{j=1}^{p} \beta_j^2 + b_\lambda \right).$$

3.1.2 Bayesian Lasso

The full conditionals for the variance parameters $\tau_j^2, j = 1,...p$, and the shrinkage parameter λ^2 are known densities so that updates for the Markov chain are available via Gibbs steps. The full conditionals for the variance parameters are

$$\pi\left(\tau_j^2 \mid \cdot\right) \propto \sqrt{\frac{1}{\tau_j^2}} \exp\left\{-\frac{\lambda^2\tau_j^2}{2m_j^2}\left(\frac{1}{\tau_j^2}-m_j\right)^2\right\},$$

where $m_j = \sqrt{\lambda^2 / \beta_j^2}$. Application of the density transformation theorem for $\omega_j{}^2 = 1/\tau_j^2$ shows that the full conditional is an inverse Gaussian distribution

$$\frac{1}{\tau_j^2} \mid \cdot \sim \text{InvGauss}\left(\frac{\sqrt{\lambda^2}}{\left|\beta_j\right|}, \lambda^2\right) \quad , j = 1,.., p\,.$$

The full conditional for the quadratic lasso parameter is given as

$$\pi\left(\lambda^2 \mid \cdot\right) \propto \left(\lambda^2\right)^{p+a_\lambda-1} \exp\left\{-\left(\frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + b_\lambda\right)\lambda^2\right\},$$

that is

$$\lambda^2 \mid \cdot \sim \text{Gamma}\left(p + a_\lambda, \frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + b_\lambda\right).$$

3.1.3 Bayesian NMIG

The diagonal matrix D_τ now contains the diagonal elements $\tau_j^2 = I_j\psi_j^2$. The full conditionals for the binary variables indicator variables are

$$\pi\left(I_j \mid y, \theta_{-I_j}\right) \propto \underbrace{(1-\omega)\frac{1}{\sqrt{\nu_0}}\exp\left\{-\frac{1}{2\nu_0\psi_j^2}\beta_j^2\right\}}_{=:A_{I,j}}\delta_{\nu_o}\left(I_j\right) + \underbrace{\omega\frac{1}{\sqrt{\nu_1}}\exp\left\{-\frac{1}{2\nu_1\psi_j^2}\beta_j^2\right\}}_{=:B_{I,j}}\delta_{\nu_1}\left(I_j\right)$$

Thus

$$\pi\left(I_j \mid y, \theta_{-I_j}\right) = \frac{1}{1 + B_{I,j}/A_{I,j}}\delta_{\nu_o}\left(I_j\right) + \frac{1}{1 + A_{I,j}/B_{I,j}}\delta_{\nu_1}\left(I_j\right)$$

with

$$\frac{A_{I,j}}{B_{I,j}} = \frac{(1-\omega)}{\omega}\frac{\sqrt{\nu_1}}{\sqrt{\nu_0}}\exp\left\{-\frac{1}{2\nu_0\psi_j^2}\beta_j^2 + \frac{1}{2\nu_1\psi_j^2}\beta_j^2\right\}$$

The full conditionals for the second variance parameter component ψ_j^2 are inverse gamma densities

$$\pi\left(\psi_j^2 \mid \cdot\right) \propto \left(\frac{1}{\psi_j^2}\right)^{\frac{1}{2}+a_\psi+1} \exp\left(-\left[b_\psi + \frac{\beta_j^2}{2I_j}\right]\frac{1}{\psi_j^2}\right),$$

so that

$$\psi_j^2 \mid \cdot \sim \text{InvGamma}\left(a_\psi + \frac{1}{2}, b_\psi + \frac{\beta_j^2}{2I_j}\right) \quad , j = 1,.., p.$$

The full conditional for the mixing parameter is a Beta density

$$\pi(\omega \mid \cdot) \propto \prod_{j=1}^{p}\left[(1-\omega)\delta_{\nu_o}\left(I_j\right) + \omega\delta_{\nu_1}\left(I_j\right)\right]1_{[0,1]}(\omega) = (1-\omega)^{n_0}\,\omega^{n_1}\,1_{[0,1]}(\omega)$$

with $n_0 := \#\left\{j : I_j = \nu_0\right\}$, $n_1 := \#\left\{j : I_j = \nu_1\right\}$. Thus

$$\omega \mid \cdot \sim \text{Beta}\left(1 + n_1; 1 + n_0\right).$$

If we use a beta prior we get $\omega \mid \cdot \sim \text{Beta}\left(a_\omega + n_1; b_\omega + n_0\right)$.

3.2 MCMC for jointly shrinking and smoothing

For the extended model (2.3), additional MH steps for drawing from full conditionals for $\zeta, \gamma_1, ..., \gamma_m$ are required. Unpenalized effects ζ are updated again by constructing a Gaussian distribution as proposal for a new state of the chain via second order Taylor expansion. Proceeding as in Section 3.1, the Gaussian proposal density $\pi(\zeta \mid \cdot)$ has precision and mean

$$\Sigma_\zeta = U'\text{diag}\left(\Lambda\left(t \mid \zeta^{(c)}, \beta, \gamma\right)\right)U + A^{-1}.$$

$$\mu_\zeta = \Sigma_\zeta^{-1}\left[U'd - U'\Lambda\left(t \mid \zeta^{(c)}, \beta, \gamma\right) + U'\text{diag}\left(\Lambda\left(t \mid \zeta^{(c)}, \beta, \gamma\right)\right)U\zeta^{(c)}\right].$$

A new proposal $\zeta^{(p)}$ of the Markov chain is accepted with probability

$$\alpha\left(\zeta^{(p)}, \zeta^{(c)}\right) = \min\left\{1, \frac{\pi\left(\zeta^{(p)} \mid \cdot\right)\hat{\pi}\left(\zeta^{(c)} \mid \zeta^{(p)}, \cdot\right)}{\pi\left(\zeta^{(c)} \mid \cdot\right)\hat{\pi}\left(\zeta^{(p)} \mid \zeta^{(c)}, \cdot\right)}\right\}.$$

For a flat prior $\pi(\zeta) \propto \text{const}$ we simply set the precision to $A^{-1} = 0$.

Basis coefficients γ_j for functions $f_j(z_j)$ are updated with IWLS proposals as described in Hennerfeind et al. (2006): Given the current state $\gamma_j^{(c)}$, proposals are drawn from a Gaussian density with precision and mean

$$\Sigma_{\gamma_j} = Z_j'\text{diag}\left(\Lambda\left(t\,|\,...,\gamma_j^{(c)},...\right)\right)Z_j + \frac{1}{\tau_j^2}K_j$$

$$\mu_{\gamma_j} = \Sigma_{\gamma_j}^{-1}\left[Z_j'd - Z_j'\Lambda\left(t\,|\,...,\gamma_j^{(c)},...\right) + Z_j'\text{diag}\left(\Lambda\left(t\,|\,...,\gamma_j^{(c)},...\right)\right)Z_j\gamma_j^{(c)}\right]$$

The full conditionals for the variance parameters δ_j^2 are (proper) inverse Gamma with parameters

$$a_j' = a_j + \tfrac{1}{2}r_j, \quad b_j' = b_j + \tfrac{1}{2}\gamma_j'K_j\gamma_j.$$

For the basis coefficients of time-varying effects, it is computationally more efficient to use MH-steps with conditional prior proposals instead of IWLS proposals.

The most costly computation in running the whole MCMC samplers are the inversions of the precision matrices within the IWLS parts of the corresponding parameter vectors. To reduce the running time in the case of high dimensional parameters, a better approach is to update these parameters in blocks of smaller size than the size of the whole parameter vector.

4 Simulations

The performance of the Bayesian Ridge, Lasso and NIG priors, is compared to the frequentist Ridge, Lasso and a backward-stepwise procedure based on the AIC criterion under Cox’s proportional hazards model. The tuning parameter λ in the frequentist Lasso and Ridge regression is estimated via n-fold generalized cross validation. The analysis of the parametric and nonparametric models with P-spline and Weibull baseline was done with the free BayesX software available from http://www.stat.uni-muenchen.de/~bayesx/. The partial likelihood based models are carried out with R version 2.6.2 from http://www.r-project.org and are implemented in functions that are available from the authors by request. Frequentistic penalized Lasso and Ridge regression is done in R with the package {penalized}.

In addition to semiparametric modelling of the baseline hazard rate in the Cox model or the P-spline based approach, we consider a simple parametric Weibull model $\lambda_0(t) = \alpha t^{\alpha-1}$ for the baseline hazard as a competitor. A gamma prior is typically employed to model the prior knowledge about the shape parameter α, i. e.

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha), \quad a_\alpha > 0, b_\alpha > 0$$
$$\pi(\alpha) \propto \alpha^{a_\alpha - 1}\exp(-b_\alpha \alpha),$$

with $E(\alpha) = a_\alpha / b_\alpha$ and $\text{Var}(\alpha) = a_\alpha / b_\alpha^2$. The update of the baseline is in this case simply replaced by the update of α. With the given prior assumption, the full conditional of α is given by

$$\pi(\alpha|\cdot) \propto \exp\sum_{i=1}^{n}\left[\delta_i\left\{\log(\alpha) + (\alpha-1)\log(t_i)\right\} - t_i^{\alpha}\exp\{x_i'\beta\}\right]\cdot\alpha^{a_\alpha - 1}\exp\{-b_\alpha\alpha\}$$

To draw a new proposal $\alpha^{(p)}$ for the shape parameter we use a Gamma distribution $\hat{\pi}\left(\cdot|\alpha^{(c)}, d_\alpha\right) \sim \text{Gamma}\left(d_\alpha \alpha^{(c)}, d_\alpha\right)$ based on the current value $\alpha^{(c)}$ which leads to the acceptance probability

$$\alpha\left(\alpha^{(p)}, \alpha^{(c)}\right) = \min\left\{1, \frac{\pi\left(\alpha^{(p)}|\cdot\right)\hat{\pi}\left(\alpha^{(c)}|\alpha^{(p)}, \cdot\right)}{\pi\left(\alpha^{(c)}|\cdot\right)\hat{\pi}\left(\alpha^{(p)}|\alpha^{(c)}, \cdot\right)}\right\}.$$

The value of d_α is determined during the burn in to achieve reasonable acceptance rates.

We measure the estimation accuracy based on the mean squared errors (MSE) over $r = 50$ runs with sample size $n = 200$ if only linear effects are modelled and $n = 1000$ if nonlinear effects are included in the predictor. Let $V = (X'X)^{-1}/n$ be the population covariance matrix of the regressors, then the MSE of $\hat{\beta}$ in the r-th simulation replication is given by

$$\text{MSE}_r(\beta) = \left(\hat{\beta}_r - \beta\right)' V\left(\hat{\beta}_r - \beta\right).$$

For the log-baseline $g_0(\cdot)$ we get

$$\text{MSE}_r(g_0) = \frac{1}{n}\left(\hat{g}_{0,r} - g_0\right)'\left(\hat{g}_{0,r} - g_0\right).$$

where $\hat{g}_0$ denotes the vector of estimates of the log-baseline in r-th simulation. To compare the results of P-spline-based log-baseline estimation and the corresponding Breslow estimates from the partial likelihood approaches, we use the trapezoidal rule to compute the cumulative baseline hazard. Concordantly, if f denotes the vector of nonlinear effects of covariate x with estimate $\hat{f}_r$, the MSE is given as

$$\mathrm{MSE}_r(f) = \frac{1}{n}\left(\hat{f}_r - f\right)'\left(\hat{f}_r - f\right).$$

Additionally, we report the average number of correct and incorrect zero coefficients in the final models achieved after applying one of the hard shrinkage rules discussed in Section 2.2.5.

Simulation settings:

For our first simulations we use the configuration from Tibshirani (1997). The covariates $x_i = \left(x_{i,1},...,x_{i,9}\right)$ are randomly drawn from a multivariate Gaussian distribution with zero mean and covariance matrix chosen such that the correlation between x_j and x_k is $\mathrm{corr}\left(x_{i,j}, x_{i,k}\right) = \rho^{|j-k|}$ with $\rho = 0.5$. The survival times $T_i, i = 1,...,n$ are generated from an exponential hazard model with constant baseline hazard $\lambda_0(t) = 1$, i. e. $\lambda(t) = \exp(x'\beta)$.

The censoring variables $C_i, i = 1,...,n$ are generated as i.i.d. draws from the uniform distribution $U\left[0, c_0\right]$ with c_0 chosen to obtain censoring rates about 25 % in each dataset. For the models

$$\begin{aligned}
&\text{Model 1:} \quad \beta = \left(-0.7,-0.7,0,0,0,-0.7,0,0,0\right)\\
&\text{Model 2:} \quad \beta = \left(-0.1,-0.1,-0.1,-0.1,-0.1,-0.1,-0.1,-0.1,-0.1\right)\\
&\text{Model 3:} \quad \beta = \left(-0.4,-0.3,0,0,0,-0.2,0,0,0\right)
\end{aligned}$$

we simulated $r = 50$ datasets with $n = 200$ life times. The first and the second model were used in Tibshirani (1997) and the third is used in Zhang and Lu (2007). For the Bayesian MCMC methods, we use 10000 iterations with a burnin of 2000 and thin the chain by 8 which results in an MCMC sample of size 1000. The hyperparameters of the Bayesian Lasso and Ridge are set to the low informative values $a_\lambda = b_\lambda = 0.01$ to allow for a greater amount of adaptiveness for the shrinkage parameter depending on the data. The hyperparameters of the Bayesian NMIG are $\nu_1 = 1$, $\nu_0 = 0.000025$, $a_\psi = 5$ and $b_\psi = 25$. These values are chosen to assign a marginal prior probability of about 0.8 to fall into the interval $[-2, 2]$ to each regression coefficient.

With the subsequent simulations we explore the changes caused by inclusion of a nonlinear effect and more flexible baseline hazards. The settings are similar to those in Hennerfeind et al. (2006). Again we consider $r = 50$ datasets but now with an increased sample size of $n = 1000$ life times. The covariates are generated independently as random draws from a uniform $U[-3,3]$ distribution. The lifetimes are generated via the inversion method (Bender, Augustin & Blettner, 2005) from the model

$$\begin{aligned}\text{Model 4}: \quad & \beta = \left(-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0\right) \\ & \lambda\left(t\right) = \lambda_0\left(t\right)\exp\left(x'\beta + \sin\left(x_{10}\right)\right).\end{aligned}$$

The covariates of the linear effects are standardized and the nonlinear effect is centered due to identification arguments leading to an intercept term in the predictor. To model more flexible baseline hazards a linear but non-Weibull baseline hazard of the form

$$\text{Model 4.a}: \quad \lambda_0\left(t\right) = 0.25 + 2t$$

and a bathtub-shaped baseline hazard

$$\text{Model 4.b}: \quad \lambda_0\left(t\right) = \begin{cases} 0.75\left(\cos\left(t\right) + 1.5\right), & t \le 2\pi \\ 0.75\left(1 + 1.5\right), & t > 2\pi \end{cases}$$

have been chosen. The latter assumes an initially high baseline risk that decreases after some time and increases again later on until time $t = 2\pi$ from where the hazard stays constant. Censoring times are generated in two steps. First a random proportion of 17% of the generated observations T_i is assigned to be censored. Then in the second step the censoring times for this random selection are drawn from the corresponding uniform distributions $U\left[0, T_i\right]$. The hyperparameters for the Bayesian penalties are not changed but here we use 20000 iterations with a burnin of 5000 and thin the chain by 10 which results in an MCMC sample of size 1500. Before we describe our results, we introduce some abbreviations to reduce the writing.

B, BT, BL, BN, BR: Bayesian models based on the full likelihood with P-spline baseline hazard without penalization (B), with the predictor that contains only the nonzero effects (BT), with Lasso (BL), NMIG (BN) and Ridge (BR) prior

SURV, SURV.T: Maximum partial likelihood of the full model and the model with the predictor that contains only the nonzero effects

STEP: Partial likelihood model with backward stepwise selection based on AIC

Pen.L, Pen.R: Penalized partial likelihood model with Lasso and Ridge penalty

PL.B, PL.BL, PL.BN, PL.BR: Bayesian models based on partial likelihood without penalization (PL.B), with Lasso (PL.BL), NMIG (PL.BN) and Ridge (PL.BR) penalty

WB.B, WB.L, WB.N, WB.R: Bayesian models based on the full likelihood with Weibull baseline hazard without penalization (WB.B), with Lasso (WB.BL), NMIG (WB.BN) and Ridge (WB.BR) penalty

For the Bayesian approaches, the hard shrinkage methods described in Section 2.2.5 are additionally assigned with

HS.CI95: if hard shrinkage is done via the 95% credible region,
HS.STD: if hard shrinkage is done via the one standard error region,
HS.IND: if hard shrinkage for NMIG is done via indicator variables.

For example, WB.N-HS.IND denotes the Bayesian Weibull model under NMIG penalty when the covariate specific indicators are used to select the covariates for the final model.

4.1 Results for model 1

Figure 4.1 shows the box plots of the mean squared errors for the regression coefficients. The Bayesian NMIG performs best within each Bayesian group of models and outperforms the STEP procedure as well as the frequentist and the Bayesian Lasso. The MSEs of the Bayesian NIG are close to the MSE of the maximum partial likelihood estimate if the predictor with the true covariate structure is used (SURV.T). The MSEs of the corresponding unpenalized Bayesian methods are comparable to the MSE of SURV.T and are omitted in the figure. The plots of Figure 4.2 display separate box plots for the corresponding estimated coefficients (except those of STEP.T, which are similar to those of SURV for the nonzero effects). If we focus on the Bayesian methods, we see that the estimates of the NMIG for the truly non-influential coefficients are much more concentrated around zero similar to STEP and Pen.L.

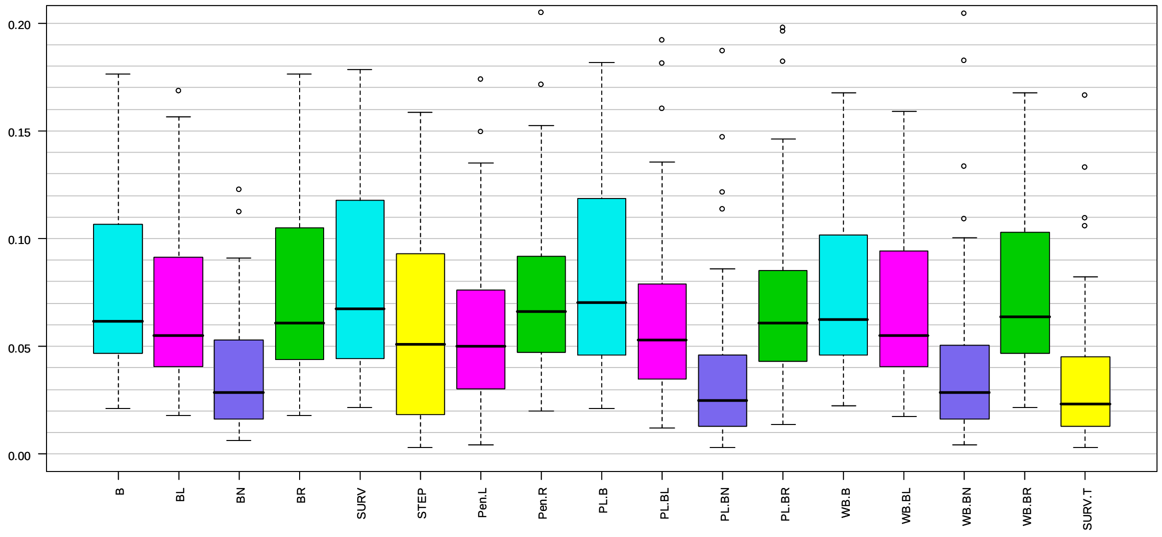

Figure 4.1: Box plots of the mean squared errors for the regression coefficients of model 1. The right box (SURV.T) shows the MSE for the maximum partial likelihood estimations when the true predictor structure is used.

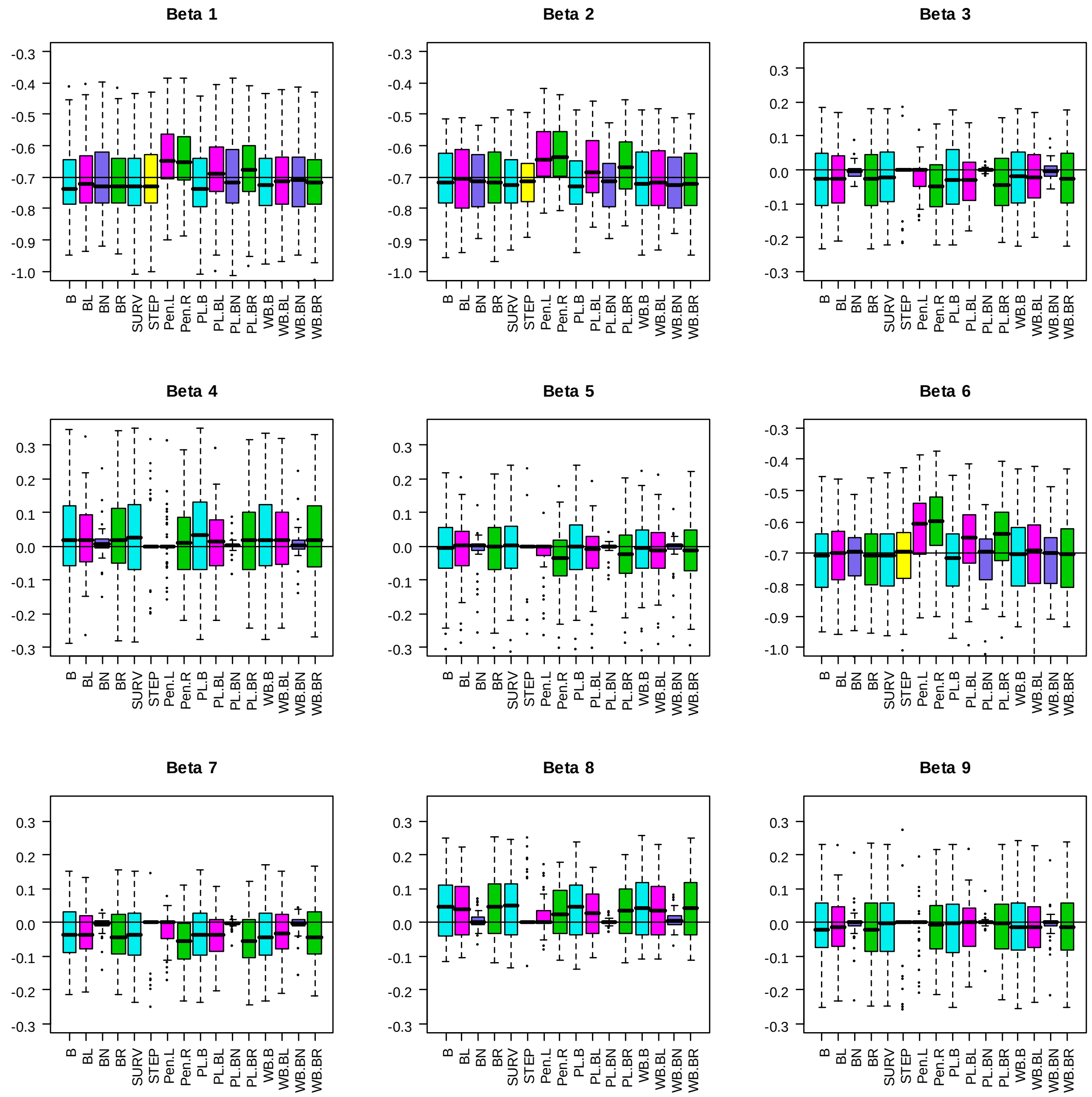

Figure 4.2: Box plot of the estimated coefficients for the simulation under model 1. The black horizontal lines mark the values of the true effects.

If we take a look at Figure 4.3, we see the different amount of penalization displayed through the box plots of the logarithm of the variance parameters τ^2 for the Bayesian Lasso, Ridge and NMIG demonstrated on the basis of the partial likelihood. For the nonzero effects $\beta_1, \beta_2, \beta_6$ the corresponding variance parameters induced by the NMIG penalty take much larger values than those of the Bayesian Lasso which results in less penalization $(1/\tau^2)$ of the nonzero coefficients estimates. The variance parameters of the zero effects are closely comparable. The tendency of small penalization for nonzero effects and larger penalization of zero effects are reflected by both, the Bayesian Lasso and the Bayesian NMIG, but the Bayesian NMIG differences for zero and nonzero effects are larger as those of the Bayesian Lasso.

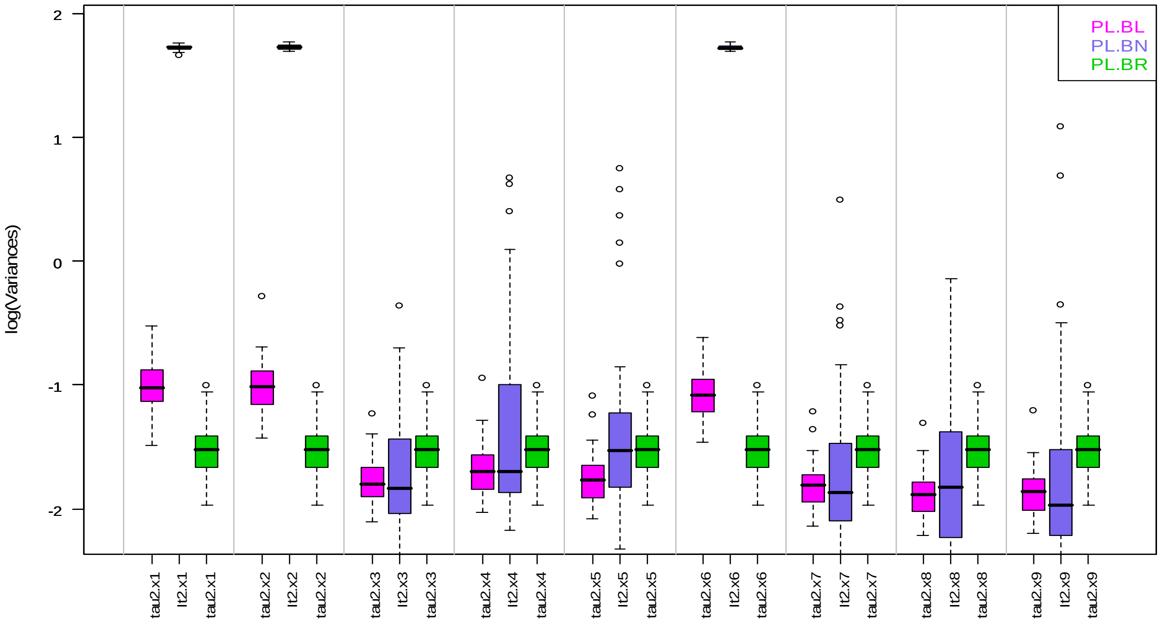

Figure 4.3: Box plot of the variance parameters for the Lasso, NMIG and Ridge penalty in the partial likelihood based Bayesian models for model 1.

The variable selection feature of the Bayesian NMIG is highlighted in Figure 4.4 where the box plots of the relative frequencies of the indicator variables $\nu_1 = 1$ are shown for the partial likelihood and full likelihood with P-spline baseline. The relative frequencies of the nonzero effects are nearly one with very small standard deviation. For the zero effects, the relative frequencies are shifted towards zero and clearly fall below the selection cut off value of 0.5. Although the frequencies of the full likelihood approach for the zero effects tend to be a little bit higher than those for the partial likelihood, the relative frequencies of the indicator variables seem to provide a good resource to select the important covariates in both cases.

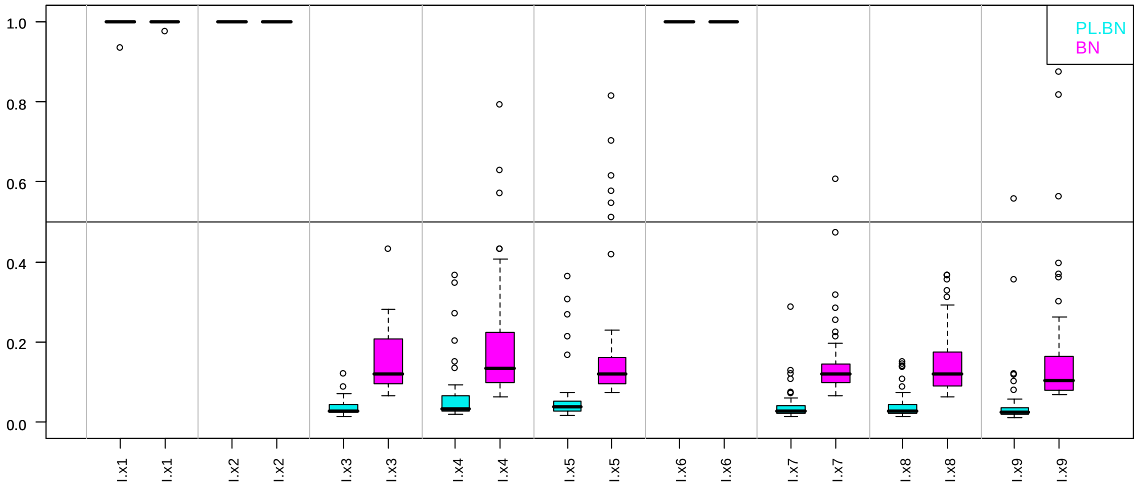

Figure 4.4: Box plot of the relative frequencies of the indicator variables for NMIG penalty if the Bayesian partial likelihood (cyan) and the full likelihood (magenta) with P-spline baseline are used for model 1.

In Figure 4.5, the box plots of the MSE for the regression coefficients are displayed together with the MSE obtained after applying the hard shrinkage criteria for the Bayesian Lasso. The results for the HS.STD criterion are omitted in this figure since they perform worse than the HS.CI95 criterion. The MSE of the Bayesian Lasso tends to be improved, if the hard shrinkage criterion is applied but the performance of the Bayesian NMIG still remains better. Furthermore, the HS.IND criterion only slightly improves the MSE of the Bayesian NMIG since the estimates of the zero effects are very close to zero anyway, i. e. it is negligible if they are removed from the final model. If we take a look on the frequencies of the selected final models, listed for the different hard shrinkage rules in Table 4.1 in column four, we find out, that all hard shrinkage criteria applied to the Bayesian NMIG for all model classes lead to the highest frequencies of the final models with the true nonzero coefficients. The best case of 50 is reached if HS.CI95 is applied, which is here not surprising because the Bayesian NMIG behaves very similar to SURV.T with asymptotic normal estimates. Additionally, the second and third column in Table 4.1 show the average number of the true estimated nonzero coefficients and true estimated zero coefficients for the 50 datasets. All hard shrinkage methods reach the optimal value of three for the true nonzero coefficients. The highest values for the true zero coefficients are again achieved for the Bayesian NMIG.

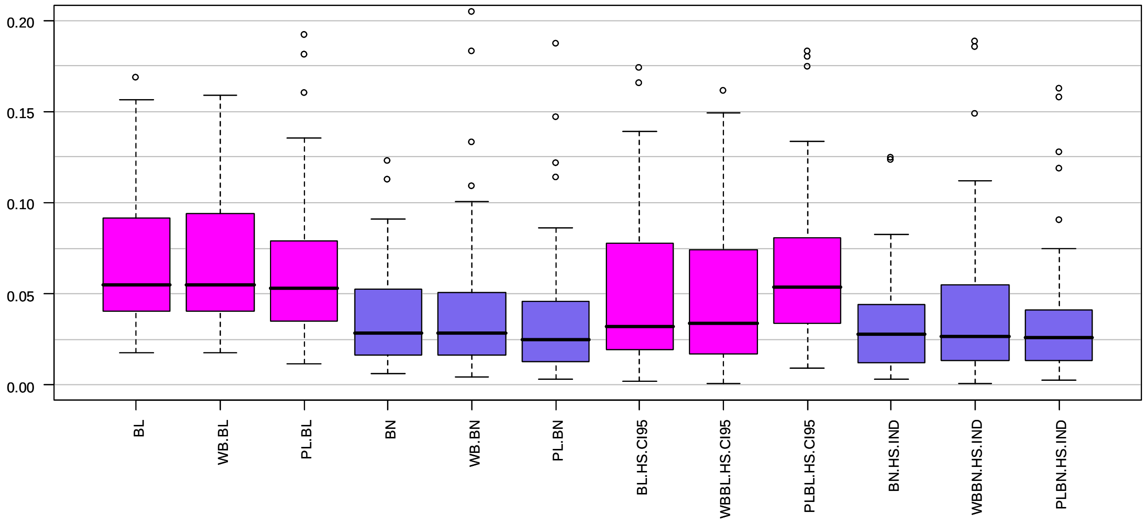

Figure 4.5: Box plot of the Bayesian Lasso (magenta) and NMIG (blue) mean squared errors for the regression coefficients of model 1. The first six boxes compare the methods if inference is based on full likelihood or partial likelihood without hard shrinkage and the last six boxes do the same if additional hard shrinkage is done for the same models.

Finally, we take a look at the MSEs of the estimates of the baselines (Figure 4.6) and the cumulative baselines (Figure 4.7) of the different model classes. For the nonparametric baseline modelled via Pspines and the parametric Weibull model in Figure 4.6 shows that the performance of the Weibull model is better than those of the nonparametric model, which is plausible since the true exponential baseline is a special Weibull baseline with $\alpha = 1$. A view at the cumulative baselines in Figure 4.7 reflects this result again and shows the outperformance of the baseline estimation produced by the full likelihood approaches. In Figure 4.8 and Figure 4.9, we see the estimates of the baseline and cumulative baseline for one selected dataset if Bayesian NMIG is applied. The vertical lines at the time axis mark the observed events. In the interval $[0,1]$, where most of the observations occur, the P-spline baseline approximates the true baseline very well. The deviations get larger, the less observations are available when time increases, which results in an increasing MSE. If we restrict the calculations of the MSEs to the interval $[0,1]$ the MSEs of the baselines as well as the MSE of the cumulative baselines of all models are similar.

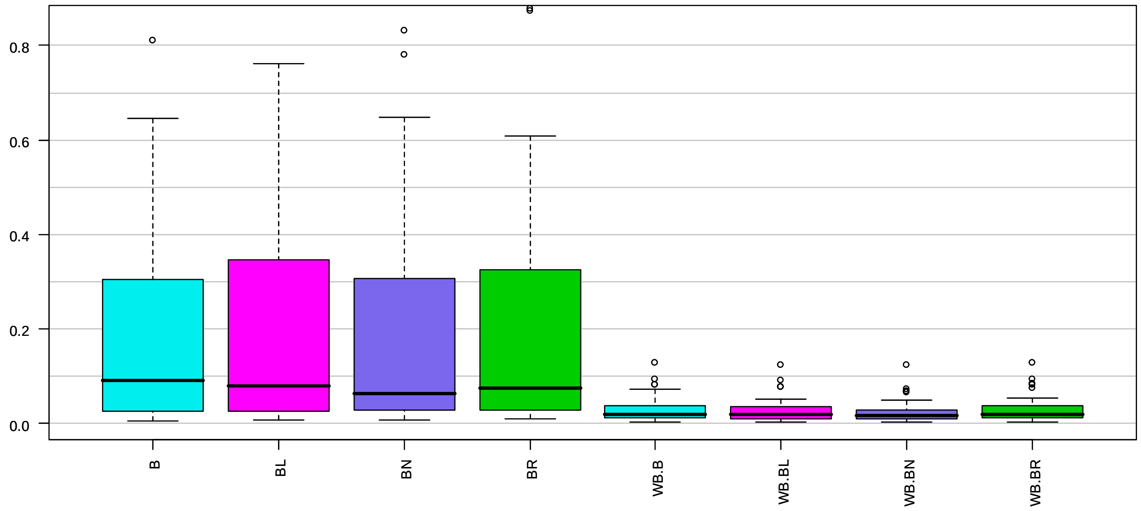

Figure 4.6: Box plot of the mean squared errors for the baseline hazards of the full likelihood based models for model 1. The first four boxes result if the baseline is modelled as P-spline, the last four boxes for the Weibull baseline.

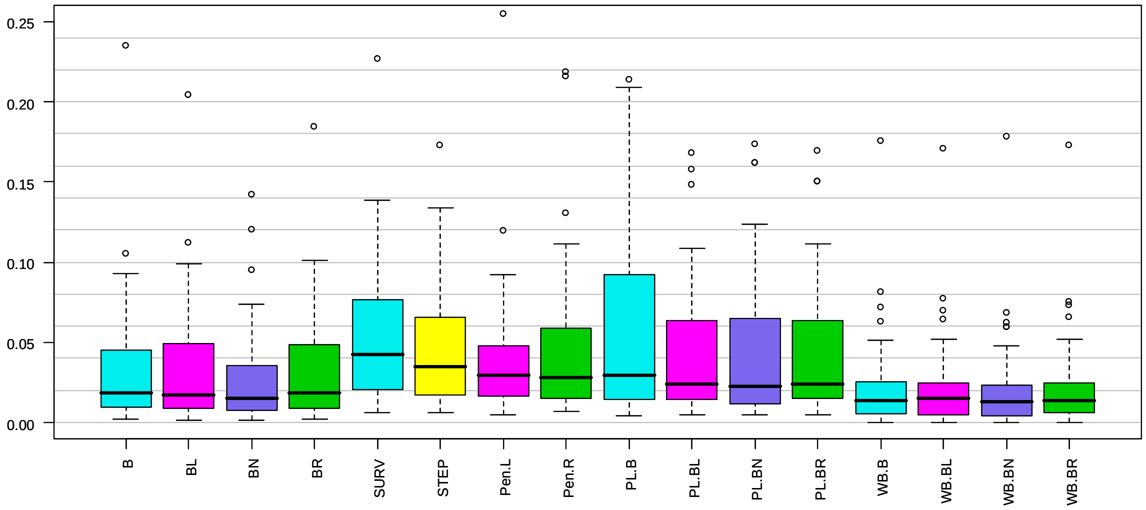

Figure 4.7: Box plot of the mean squared errors for the cumulative baseline hazards of model 1.

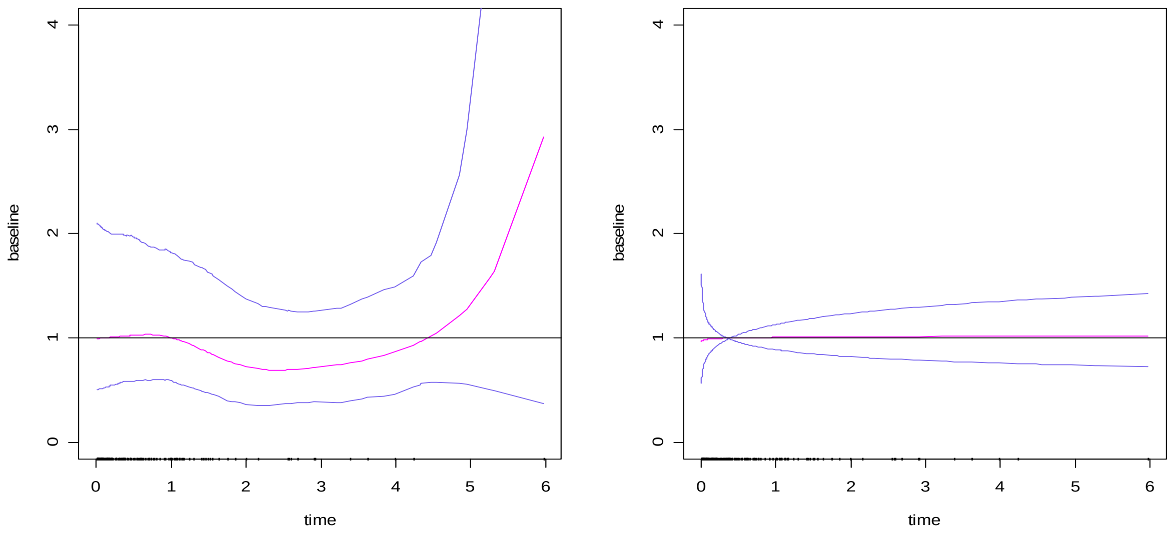

Figure 4.8: Estimation of the baseline hazard (magenta lines) for one selected dataset under model 1 together with the 2.5% and 97.5% empirical quantiles (blue lines). On the left side the baseline is modelled via P-spline, on the right side the Weibull baseline is used. The black lines mark the true exponential baseline and the vertical lines at the time axis mark the observed events.

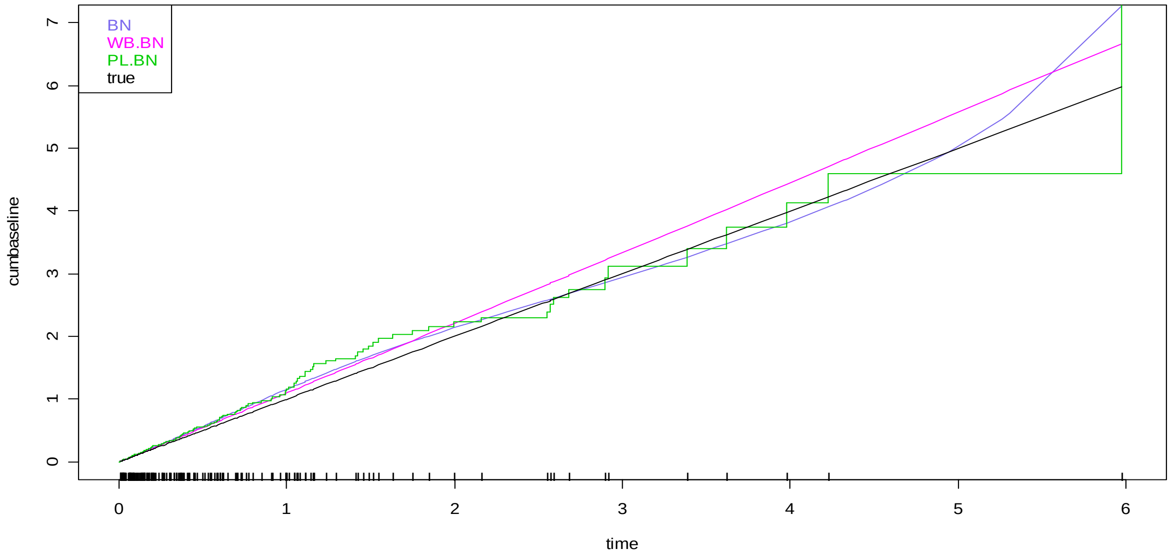

Figure 4.9: Cumulative baseline hazards under the Bayesian NIG penalty for one selected dataset under model 1. The black line marks the true exponential cumulative baseline, the vertical lines at the time axis mark the observed events.

4.2 Results for model 2

Figure 4.10 shows the MSEs for the different methods if all nine covariates in the predictor are assigned small but nonzero effects. The MSEs for the Bayesian NIG method based on the full likelihood are comparable to the MSE of the stepwise procedure STEP and are larger then the MSEs of the Bayesian Lasso, Ridge and the unpenalized Bayesian approach. The increase of the MSE can be explained when taking a look at Figure 4.12, where the relative frequencies of the indicator variables produced by the Bayesian NMIG penalty are displayed. Almost all relative frequencies of the indicators are of comparable size and tend to be close to zero, which results in a heavy penalization of all covariates and accordingly in point estimates for the regression coefficients that are very close to zero and don't reflect the true model. The stepwise procedure behaves similar by producing zero estimates for the small nonzero effects. The best performances are obtained for the Ridge- and Lasso-penalty based methods. It is clear, that in the setting of model 2 with small nonzero effects, the application of the hard shrinkage criteria can not improve the MSE (Figure 4.11). In Table 4.1, the number of correctly identified nonzero and zero estimated coefficients are collected for the hard shrinkage criteria. The best value of nine is reached for none of the approaches. The estimation of the baselines and cumulative baselines leads to comparable results as in model 1 and are summarized in terms of the MSE in Figure 4.13.

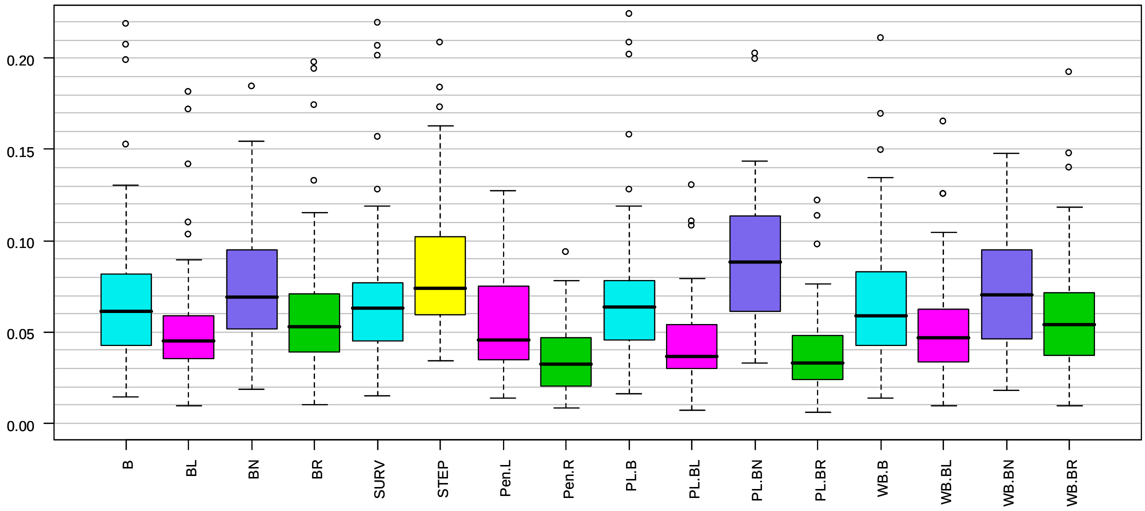

Figure 4.10: Box plot of the mean squared errors for the regression coefficients of model 2.

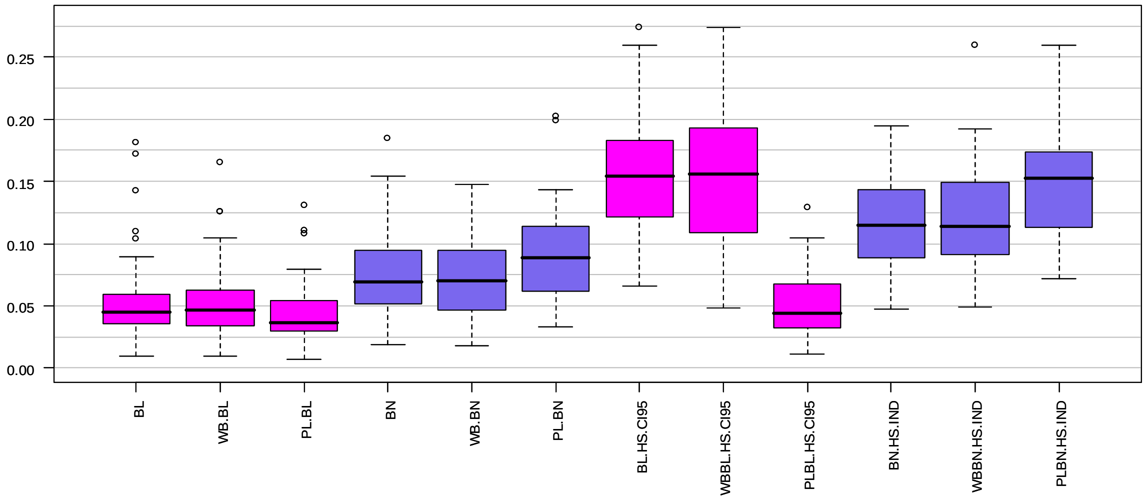

Figure 4.11: Box plot for the Bayesian Lasso (magenta) and NMIG (blue) mean squared errors for the regression coefficients of model 2. The first six boxes compare the methods if inference is based on full likelihood or partial likelihood without hard shrinkage, while the last six boxes correspond to results when additional hard shrinkage is applied to the same models.

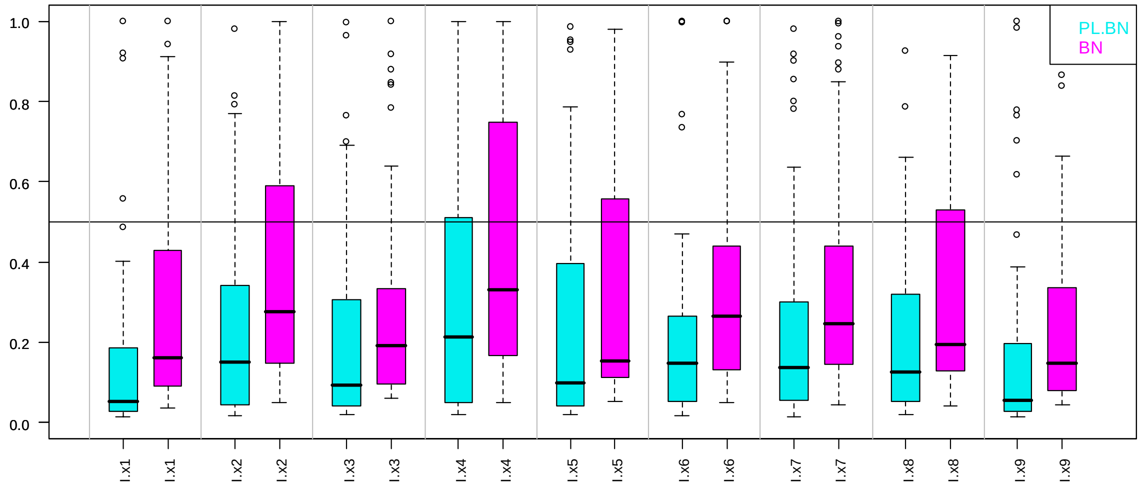

Figure 4.12: Box plot of the relative frequencies of the indicator variables for the NMIG penalty if the Bayesian partial likelihood (cyan) and the full likelihood (magenta) with P-spline baseline are used for model 2.

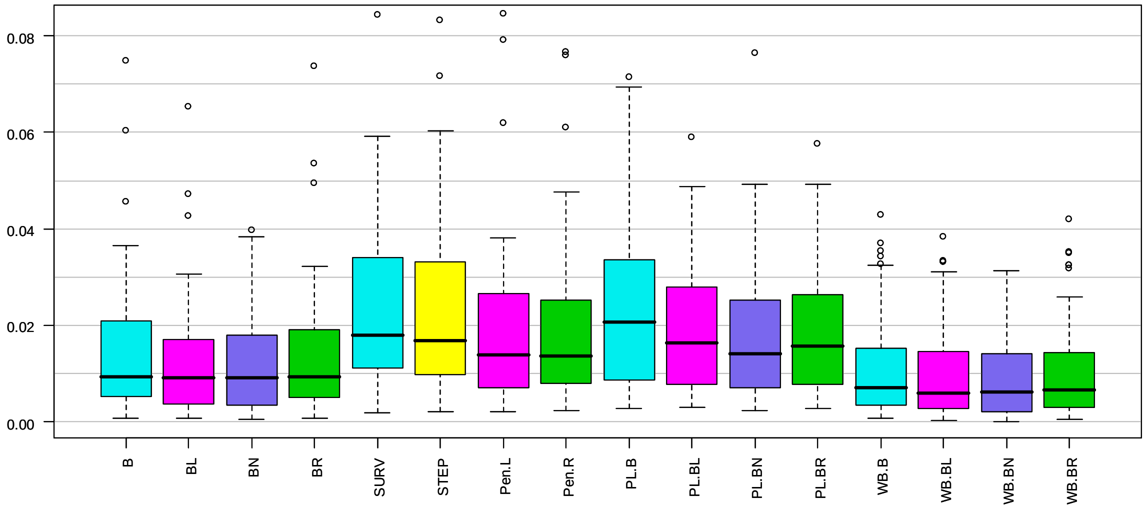

Figure 4.13: Box plot of the mean squared errors for the cumulative baseline hazards of model 2.

4.3 Results for model 3

Figure 4.14 shows the MSEs achieved in the estimation of model 3. Again, the MSEs of the maximum partial likelihood estimators for the true predictor structure are recorded as a benchmark result. The penalty based approaches achieve lower MSEs than those without penalisation and the best MSEs are derived with the Lasso penalty in combination with the partial likelihood. The performance of the full likelihood based methods are similar if the Lasso and NMIG penalty are employed. Taking a look at Figure 4.15 where the relative frequencies of the value $\nu_1 = 1$ of the Bayesian NMIG are displayed, reveals that the relative frequency for the largest effect with value -0.4 is nearly 1 and for the second largest effect with value -0.2 the cut off value of 0.5 is frequently passed over. Further simulations (not presented in this work) have shown that effects with absolute values larger than 0.3 can be separated from the zero effects very well by the indicator variables relative frequencies, if the same prior settings as noted above are used for comparable models. As in model 2, the hard shrinkage criteria do not improve the MSE (Figure 4.16). For model 3, the STEP procedure detected the true effects in most cases followed by the frequentist Lasso, compare Table 4.1.

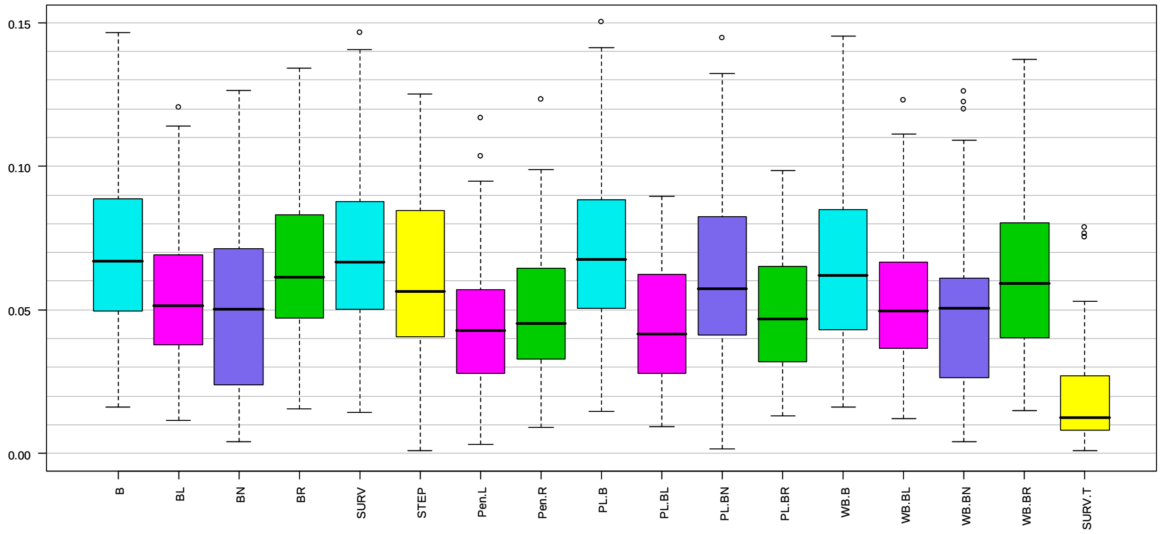

Figure 4.14: Box plot of the mean squared errors for the regression coefficients of model 3. The right box (SURV.T) shows the MSE for the maximum partial likelihood estimates when the true predictor structure is used.

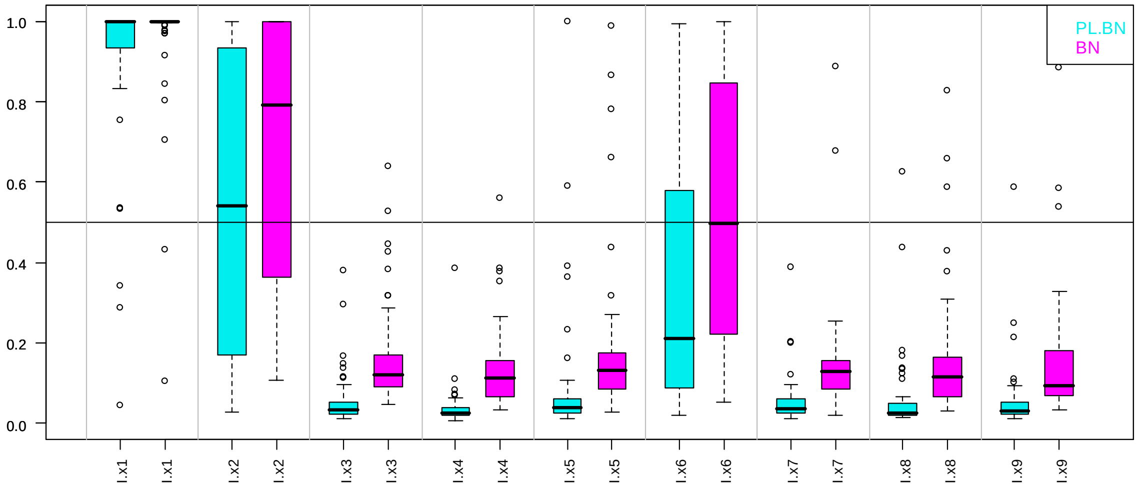

Figure 4.15: Box plot of the relative frequencies of the indicator variables for NMIG penalty if the Bayesian partial likelihood (cyan) and the full likelihood (magenta) with P-spline baseline are used for model 3.

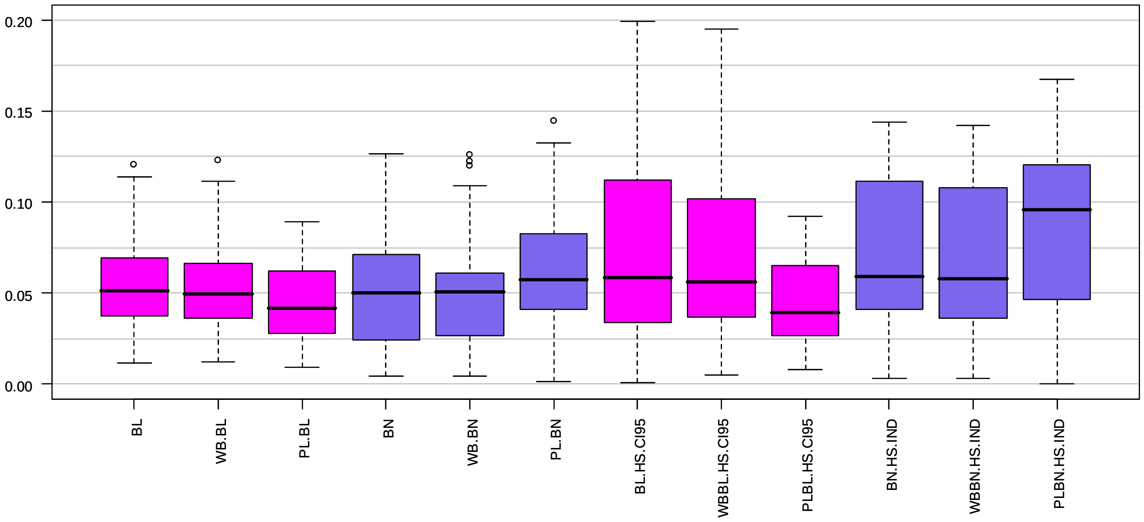

Figure 4.16: Box plot for the Bayesian Lasso (magenta) and NMIG (blue) mean squared errors for the regression coefficients of model 3. The first six boxes compare the methods if inference is based on full likelihood or partial likelihood without hard shrinkage while the last six boxes correspond to results when hard shrinkage is applied to the same models

	Model 1			Model 2			Model 3		
	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	MF	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	MF	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	MF
BEST	3	6	50	9	0	50	3	6	50
STEP	3	4.9	19	3.94	0	0	2.66	4.58	6
Pen.L	3	3.6	2	6.42	0	2	2.88	3.86	7
B.HS-STD	3	4.18	5	4.58	0	0	2.8	3.9	2
WBB.HS-STD	3	4.28	4	4.52	0	0	2.78	3.94	3
PLB.HS-STD	3	4.18	4	4.54	0	0	2.78	3.86	2
BL.HS-STD	3	4.38	6	4.38	0	0	2.78	4.26	4
WBBL.HS-STD	3	4.56	9	4.36	0	0	2.78	4.4	8
PLBL.HS-STD	3	4.7	8	4.04	0	0	2.76	4.88	9
BN.HS-STD	3	5.82	42	2.04	0	0	2.12	5.74	12
WBBN.HS-STD	3	5.82	42	1.94	0	0	2.06	5.78	14
PLBN.HS-STD	3	5.98	49	1.24	0	0	1.66	5.92	6
BR.HS-STD	3	4.3	4	4.66	0	0	2.8	3.94	2
WBBR.HS-STD	3	4.26	3	4.52	0	0	2.8	4.08	4
PLBR.HS-STD	3	4.42	6	4.74	0	0	2.76	4.5	7
B.HS-CI95	3	5.68	38	1.62	0	0	2.26	5.66	15
WBB.HS-CI95	3	5.64	36	1.48	0	0	2.22	5.66	14
PLB.HS-CI95	3	2.92	1	5.84	0	0	2.92	2.94	0
BL.HS-CI95	3	5.74	40	1.32	0	0	2.2	5.76	15
WBBL.HS-CI95	3	5.68	38	1.32	0	0	2.16	5.78	13
PLBL.HS-CI95	3	3.26	1	5.9	0	0	2.88	3.64	3
BN.HS-CI95	3	6	50	0.6	0	0	1.54	5.98	3
WBBN.HS-CI95	3	6	50	0.56	0	0	1.52	5.98	4
PLBN.HS-CI95	3	5.94	47	1.98	0	0	1.98	5.82	11
BR.HS-CI95	3	5.68	37	1.44	0	0	2.3	5.68	16
WBBR.HS-CI95	3	5.62	35	1.34	0	0	2.18	5.68	13
PLBR.HS-CI95	3	2.94	0	6.3	0	0	2.9	3.32	2
BN.HS-IND	3	5.74	40	2.3	0	0	2.14	5.7	12
WBBN.HS-IND	3	5.8	41	2.2	0	0	2.12	5.72	14
PLBN.HS-IND	3	5.98	49	1.42	0	0	1.74	5.92	8

Table 4.1: Number of “true” estimated coefficients where $\hat{\beta} \neq 0, \beta \neq 0$ denotes the case that the estimated effect is nonzero ($\hat{\beta} \neq 0$) when corresponding true effect is nonzero ($\beta \neq 0$) and $\hat{\beta} = 0, \beta = 0$ denotes the case that the estimated effect is zero ($\hat{\beta} = 0$) when corresponding true effect is zero ($\beta = 0$) The columns (MF) display the frequencies of the final models that contain only the three effects $\beta_1 \neq 0, \beta_2 \neq 0, \beta_6 \neq 0$ for model 1, 2 and 3.

4.4 Results for model 4

We only briefly summarize results for the models 4.a and 4.b due to the similarity to the results of model 1. We restrict ourselves to the Bayesian methods based on the full likelihood with P-spline approximation for the baseline. At present, there are no distributed packages in R available to perform frequentist Lasso regression in combination with nonlinear effects for Cox PH models. Ridge regression is possible but the shrinkage parameter lambda has to be prespecified.

In Figure 4.17, the MSEs of the estimated regression coefficients are shown together with the MSEs if the hard shrinkage criteria are applied. As in model 1 the Bayesian NMIG performs better than the Bayesian Lasso regardless of whether hard shrinkage is applied or not. The comparison of the variance parameters and the indicator variables are leading to similar results as in model 1.

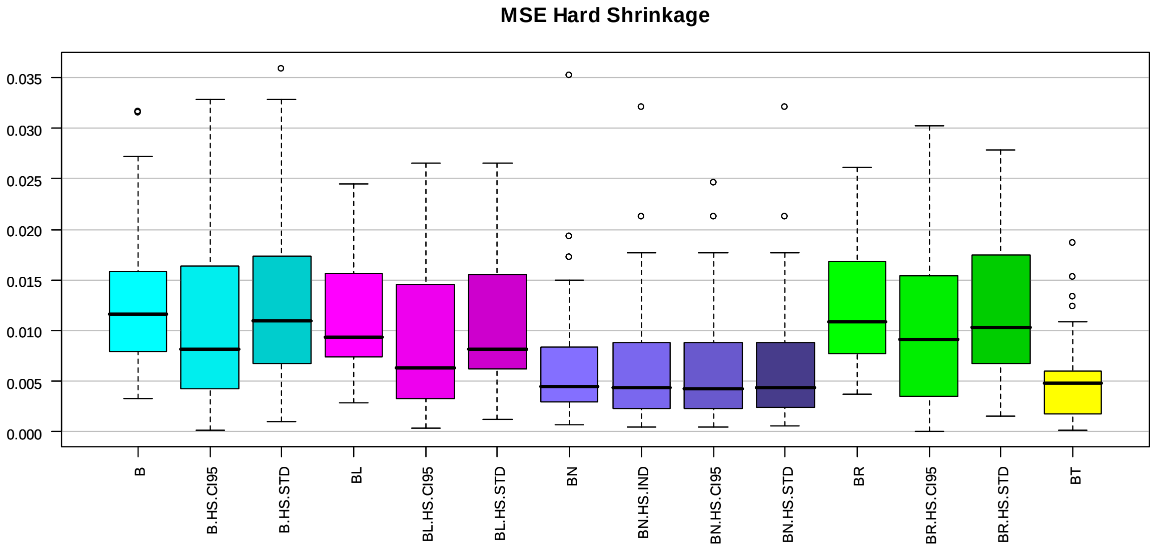

Figure 4.17: Box plot of the mean squared errors for the regression coefficients of model 4.a in combination with the mean squared errors if hard shrinkage is used.

Figure 4.18 shows the results for the baseline and cumulative baseline estimation for one selected dataset. Again, we see that the P-spline approximation of the baseline performs very well in the time region where most of the events occur. The approximation works as good as in model 1 but in a larger time interval since the sample size was increased to account for the more complicated setting. If we take a look at the Deviance Information Criterion (DIC) and the effective number of parameters (Spiegelhalter et al., 2002) that are shown in Figure 4.19, the Bayesian NIG in comparison to the Bayesian Lasso has a lower effective number of parameters but still yields a comparable DIC. The difference to model 1 is the additional inclusion of a nonlinear effect. The shrinkage methods for the linear effects do not affect the performance of the estimates of the nonlinear effect as shown in Figure 4.20, where the estimated P-spline is visualized together with the 2.5% and 97.5% empirical quantiles for one selected dataset for the model without penalization and for the model NMIG penalty. For model 4.b with bathtub shaped baseline, the box plots of the estimated coefficients are shown in Figure 4.21 The box plots for the different methods are very similar except those of the Bayesian NMIG for the zero coefficients which show a higher concentration around zero. The hard shrinkage rules are leading to comparable results since the non-influential covariates are assigned an effect very close to zero anyway.

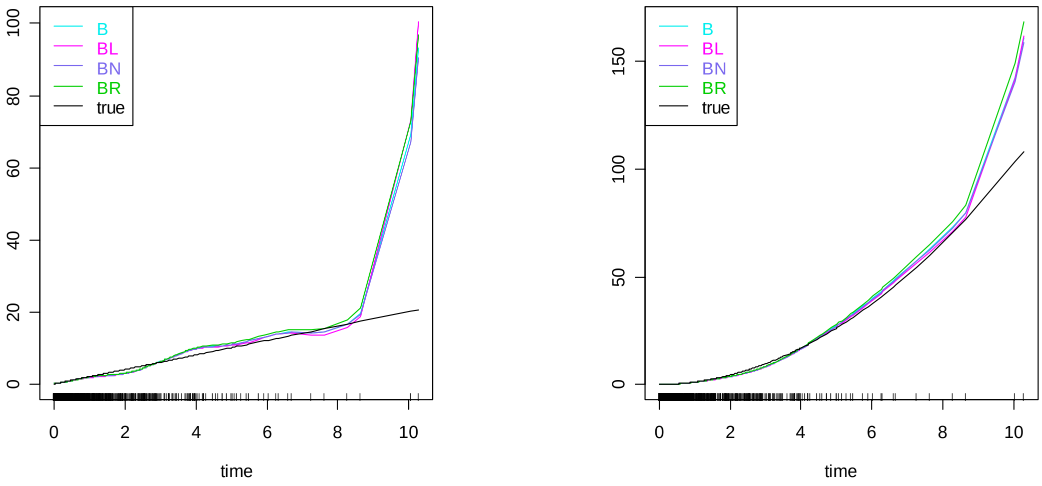

Figure 4.18: Baseline hazards (left side) and cumulative baseline hazards (right side) under the Bayesian NIG penalty for one selected dataset and model 4.a. The black lines mark the true exponential baseline and cumulative baseline.

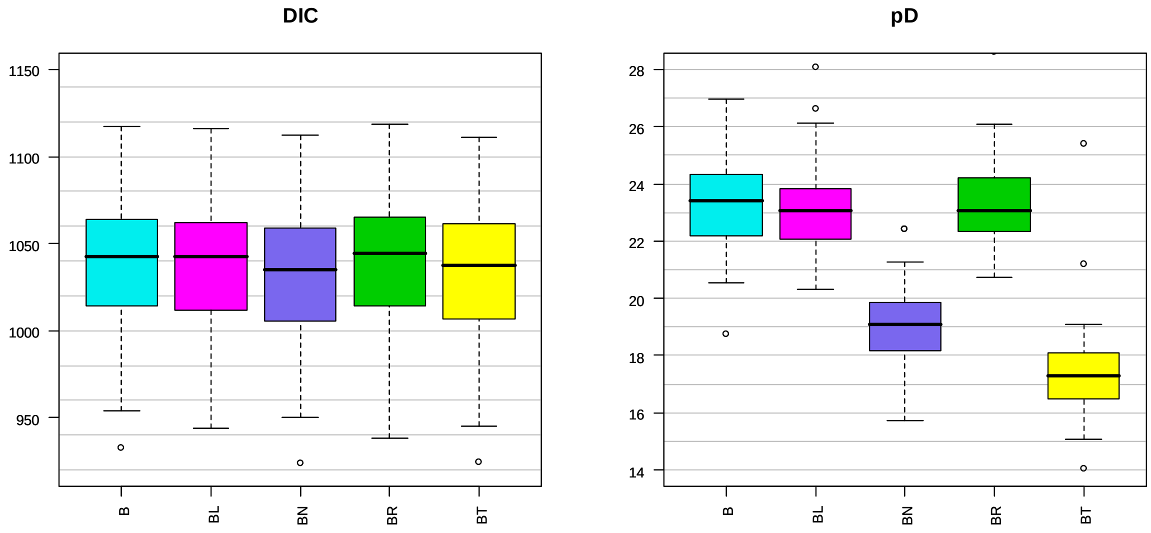

Figure 4.19: Box plot of the Deviance Information Criterion (left side) and the effective number of parameters (right side) for model 4.a.

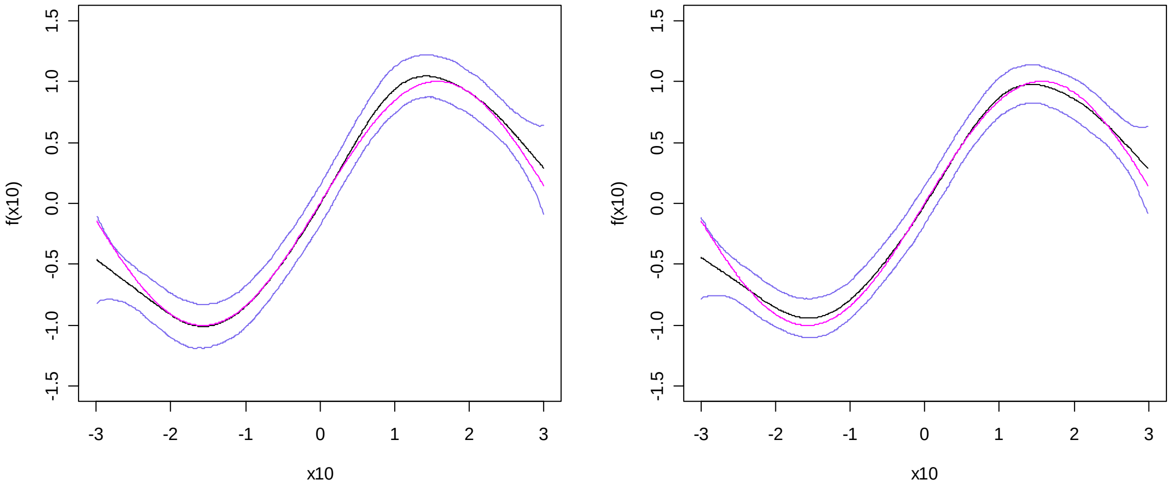

Figure 4.20: Estimation of the nonlinear effect f(x)=sin(x) (magenta lines) for 1 selected data set under model 4.a together with the 2.5% and 97.5% empirical quantiles (blue lines). On the left side the results with no penalty for the linear effects are displayed, while the right side shows results with the NMIG penalty. The black lines indicate the true effect.

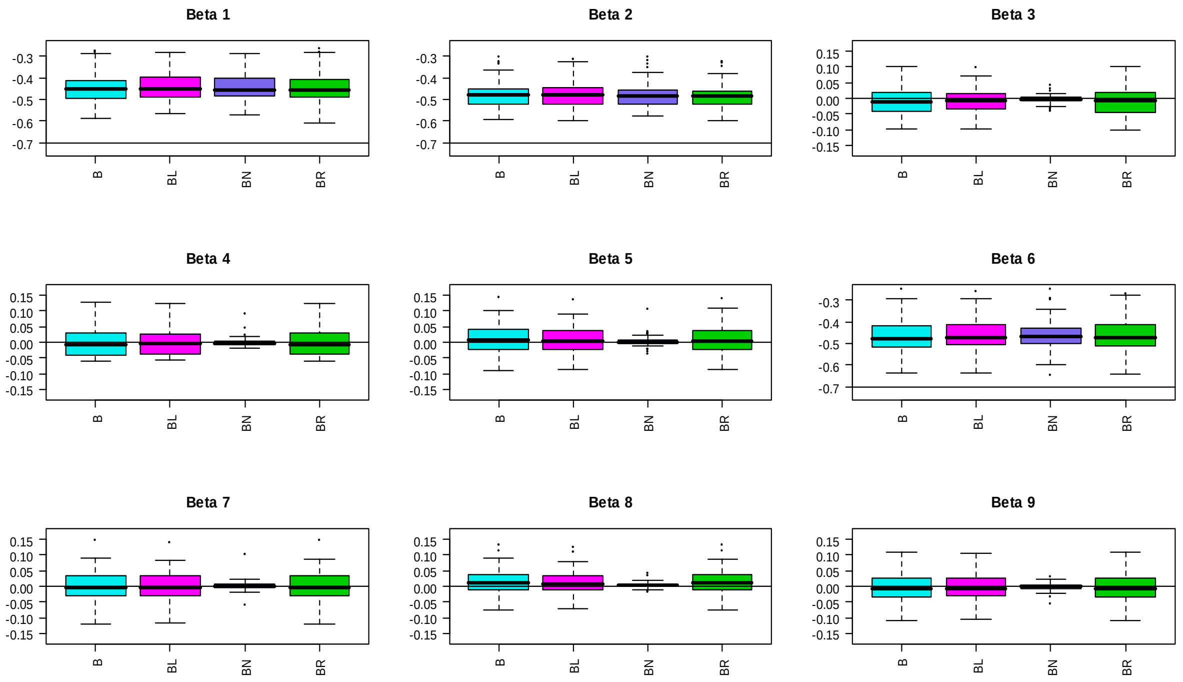

Figure 4.21: Box plot of the estimated coefficients for the simulation under model 4.b. The black horizontal lines mark the values of the true linear effects.

As we see in Figure 4.22, the relative frequencies of the indicator variables reflect the number of true nonzero effects and the number of true zero effects well. The results for baseline estimation (Figure 4.23) are as good as in model 4.a and the results of the different methods for the nonlinear effect are again comparable to each other but they do not reach the very good performance as in model 4.a. The frequencies of the final models and the number of true estimated coefficients are collected in Table 4.2 and the results are almost comparable to those of model 4.a.

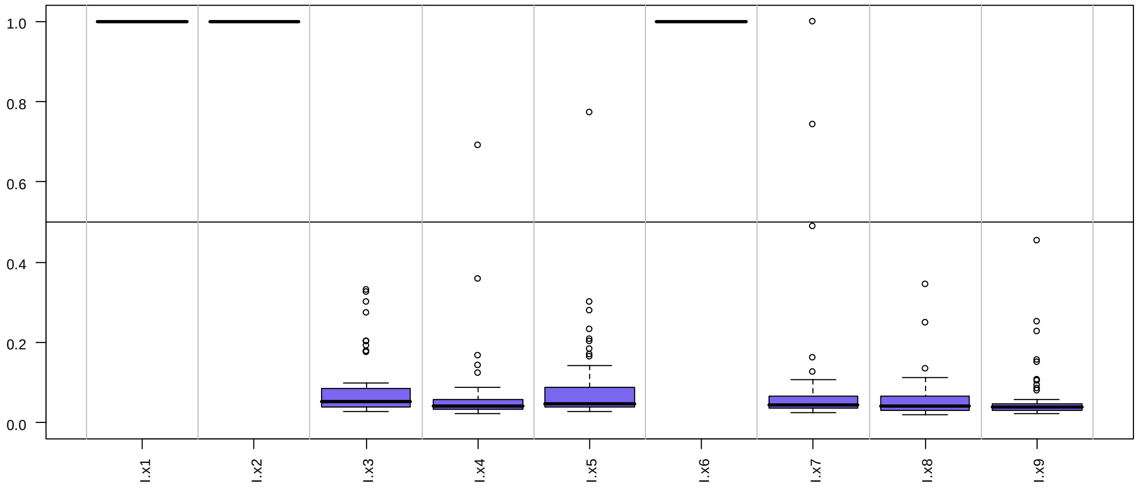

Figure 4.22: Box plots of the indicator variables for NMIG penalty in model 4.b.

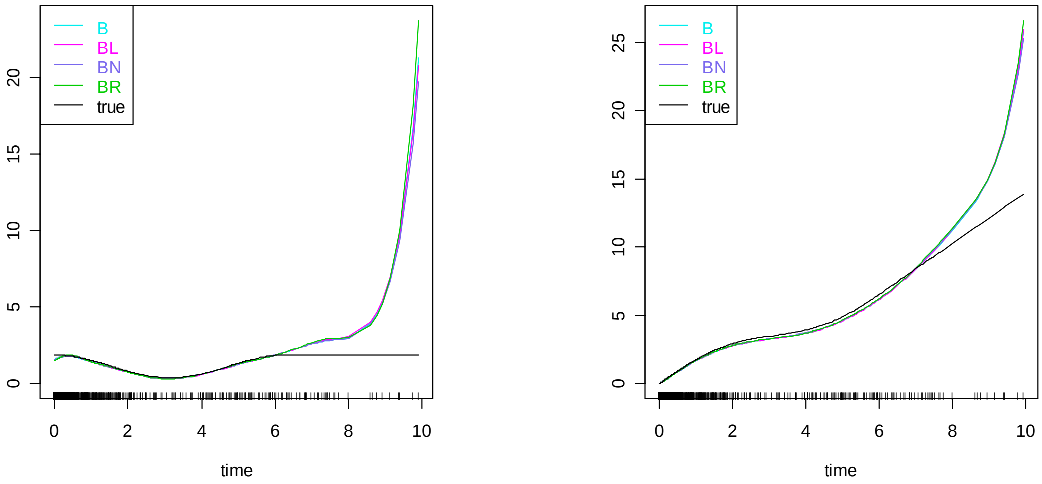

Figure 4.23: Baseline hazards (left side) and cumulative baseline hazards (right side) under the Bayesian NIG penalty for one selected dataset under model 4.b. The black lines indicate the true baseline and cumulative baseline.

	Model 4.a			Model 4.b		
	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	MF	$\hat{\beta} \neq 0$ $\beta \neq 0$	$\hat{\beta} = 0$ $\beta = 0$	MF
BEST	3	6	50	3	6	50
B.HS-STD	3	4.30	8	3	4.04	7
BL.HS-STD	3	4.30	7	3	4.32	7
BN.HS-STD	3	5.84	42	3	5.84	42
BR.HS-STD	3	4.22	9	3	4.06	6
B.HS-CI95	3	5.66	36	3	5.64	34
BL.HS-CI95	3	5.70	38	3	5.66	35
BN.HS-CI95	3	6.00	50	3	5.98	49
BR.HS-CI95	3	5.60	35	3	5.64	35
BN.HS-IND	3	5.94	47	3	5.92	46

Table 4.2: Number of “true” estimated coefficients where $\hat{\beta} \neq 0, \beta \neq 0$ denotes the case that the estimated effect is nonzero ($\hat{\beta} \neq 0$) when the corresponding true effect is nonzero ($\beta \neq 0$) and $\hat{\beta} = 0, \beta = 0$ denotes the case that the estimated effect is zero ($\hat{\beta} = 0$) when the corresponding true effect is zero ($\beta = 0$) The columns (MF) display the frequencies of the final models that contain only the three effects $\beta_1 \neq 0, \beta_2 \neq 0, \beta_6 \neq 0$ for model 4.a and 4.b.

5 Application

The presented methods are applied to the primary biliary cirrhosis (PBC) data provided for example in the R package survival or the book-homepage of Therneau and Grambsch (2000). Primary biliary cirrhosis is an autoimmune disease of the liver, marked by the slow progressive destruction of the small bile ducts (bile canaliculi) within the liver. When these ducts are damaged, bile builds up in the liver and over time damages the tissue. This can lead to scarring, fibrosis, cirrhosis, and ultimately liver failure and death of the patient. In Section 5.1 we give a description of the data and refer to Therneau and Grambsch (2000) for a more detailed account and an extended frequentist analysis. The PBC data is also used in Tibshirani (1997) to compare the variable selection property of the Lasso with a backward-forward stepwise procedure based on p-values or the BIC criterion. Further, Zhang and Lou (2007) applied their adaptive Lasso on this data. They extended the Lasso with individual weights w_i introduced in the penalization term $\sum|\beta_i|w_i$ of the regression coefficients which leads to different penalties. They state that under an appropriate choice of the weights, e.g. as the inverse maximizer of the log partial likelihood, and the shrinkage parameter the adaptive Lasso can asymptotically perform as well as if the correct submodel was known.

We compare our methods with a stepwise-backward procedure for Cox's regression model based on the AIC and frequentist Lasso regression provided in the R package {penalized} (Goeman, 2007). In the latter case, the penalization parameter was determined by n-fold generalized cross validation. An alternative implementation of Lasso regression based on the Cox model is provided in the R package {glmpath} (Park and Hastie, 2006). For the Bayesian MCMC methods, we use 20000 iterations with a burnin of 5000 and thin the chain by 10 which results in an MCMC sample of size 1500. The hyperparameters of the Bayesian Lasso and Ridge are the same as in the simulations settings of Section 4.

5.1 Primary Biliary Cirrhosis of Liver

The data has been collected from the Mayo Clinic trial in primary biliary cirrhosis of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data considered here consists of the additional 106 cases as well as the 312 randomized participants. Discarding observations with missing values leaves $n=276$ observations with 58.42 % censoring. The covariates used for analysis are

age	age in years
alb	albumin in gm/dl
alkphos	alkaline phosphatase in U/liter
ascites	presence of asictes (0 = no, 1 = yes)
bili	serum bilirubin in mg/dl
chol	serum cholesterol in mg/dl
copper	urine copper in ug/day
edtrt	presence of edema (0.0 = no edema and no diuretic therapy for edema; 0.5 = edema present without diuretics, or edema resolved by diuretics; 1.0 = edema despite diuretic therapy)
hepmeg	presence of hepatomegaly, i. e. enlarged liver (0 = no, 1 = yes)
platelet	platelets per cubic ml / 1000
protime	standardized blood clotting time, prothrombin time in seconds
sex	sex (0 = male, 1 = female)
sgot	liver enzyme SGOT in U/ml
spiders	blood vessel malformations in the skin, presence of spiders (0 = no, 1 = yes)
stage	histologic stage of disease
trig	triglicerides in mg/dl
trt	treatment/drug (1= D-penicillamine, 2 = placebo)

To make our results comparable to those in Tibshirani (1997) the covariates were standardized to have zero mean and unit variance. The point estimates together with the corresponding standard deviations for the regression coefficients are displayed in the upper plot of Figure 5.1. The lower plot shows results from the Bayesian methods after applying the hard shrinkage rules to select covariates for the final model together with the results from the stepwise procedure and the Lasso. For those coefficients that are not included in the final model the standard deviations are set to zero. The standard deviations for the Lasso are obtained by the approximate method described in Tibshirani (1997). All methods are leading to models that include the five covariates age, alb, bili, stage and copper and eliminate hepmeg, platelet, spiders, trt and trig. The covariate ascites is only chosen by the Lasso, but the effect of ascites is very small. Obviously, the NMIG methods shrink small effects to a larger extent than the Lasso based methods, so that most of the remaining covariates (except edrt) are excluded if selection is implemented based on the frequency of the NMIG indicator variables. Figure 5.2 shows the indicator frequencies of $\nu_1 = 1$ for the partial and the full likelihood.

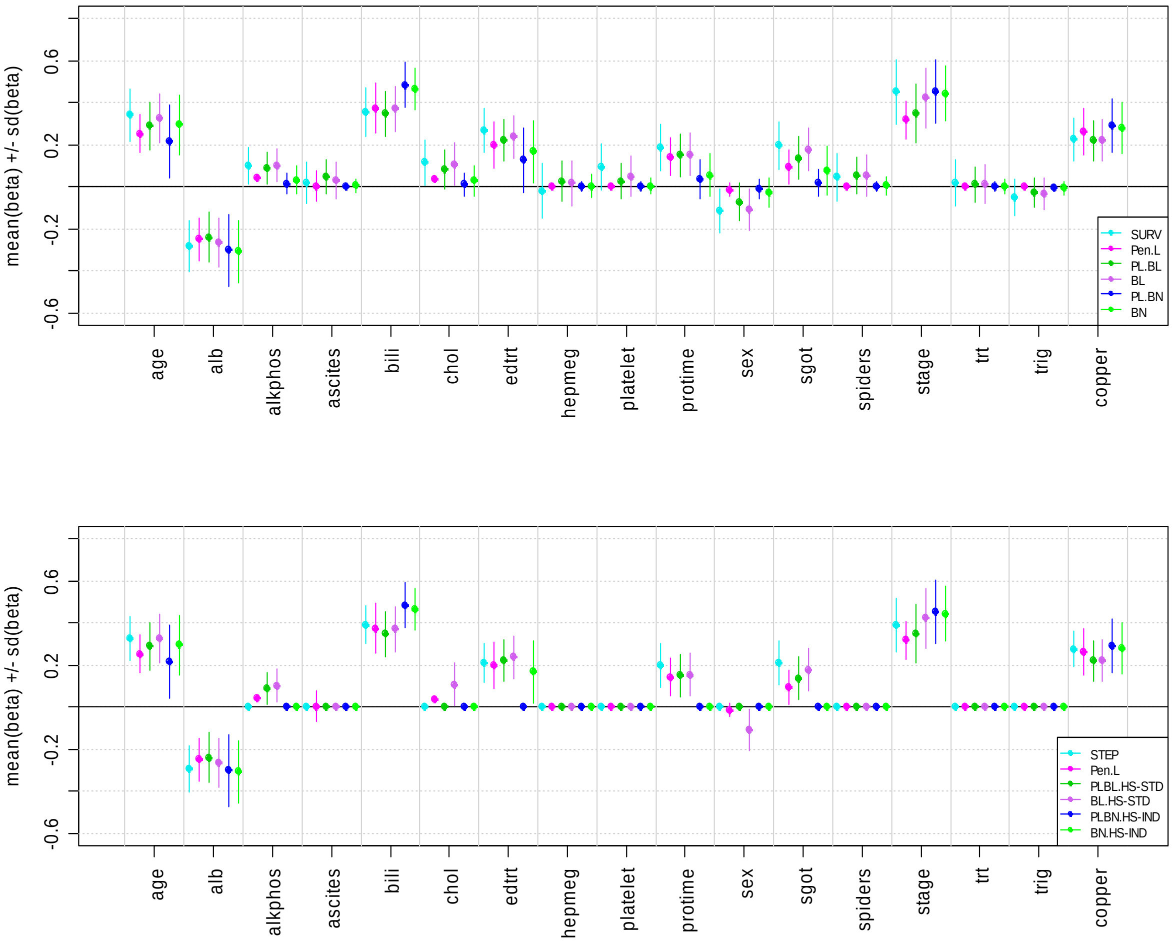

Figure 5.1 Estimated coefficients and standard errors without (upper panel) and with variable selection (lower panel). (SURV: Maximum partial likelihood, PEN.L: Lasso with package penalized, PL.BL: Bayesian Lasso based on partial likelihood, BL: Bayesian Lasso based on full likelihood, PL.BN: Bayesian NMIG based on partial likelihood, BN: Bayesian NMIG based on full likelihood. PL.BL-HS-STD: Bayesian Lasso based on partial likelihood with hard shrinkage via standard deviation, BL: Bayesian Lasso based on full likelihood with hard shrinkage via standard deviation, PL.BN: Bayesian NMIG based on partial likelihood with hard shrinkage via indicator variables, BN: Bayesian NMIG based on full likelihood with hard shrinkage via indicator variables)

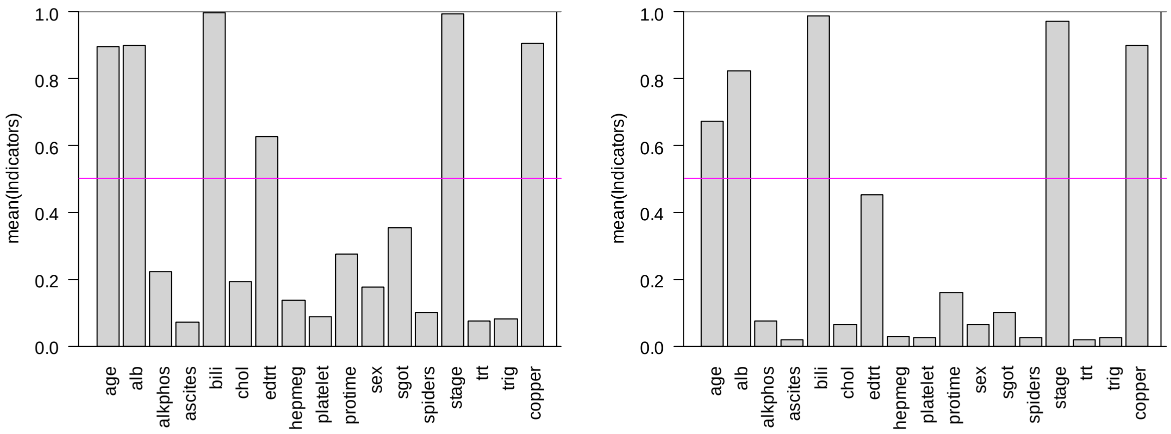

Figure 5.2 Frequencies of the of the Bayesian NMIG indicator variable value ν_1 Left side: Based on full likelihood. Right side: Based on partial likelihood.

For the Bayesian Lasso and NMIG based on the full likelihood, the log baseline hazards are depicted in Figure 5.3 .The corresponding cumulative baseline hazards (obtained by applying the trapezoidal rule for integration) are shown in Figure 5.4 together with Breslow’s estimate for the partial likelihood based methods.

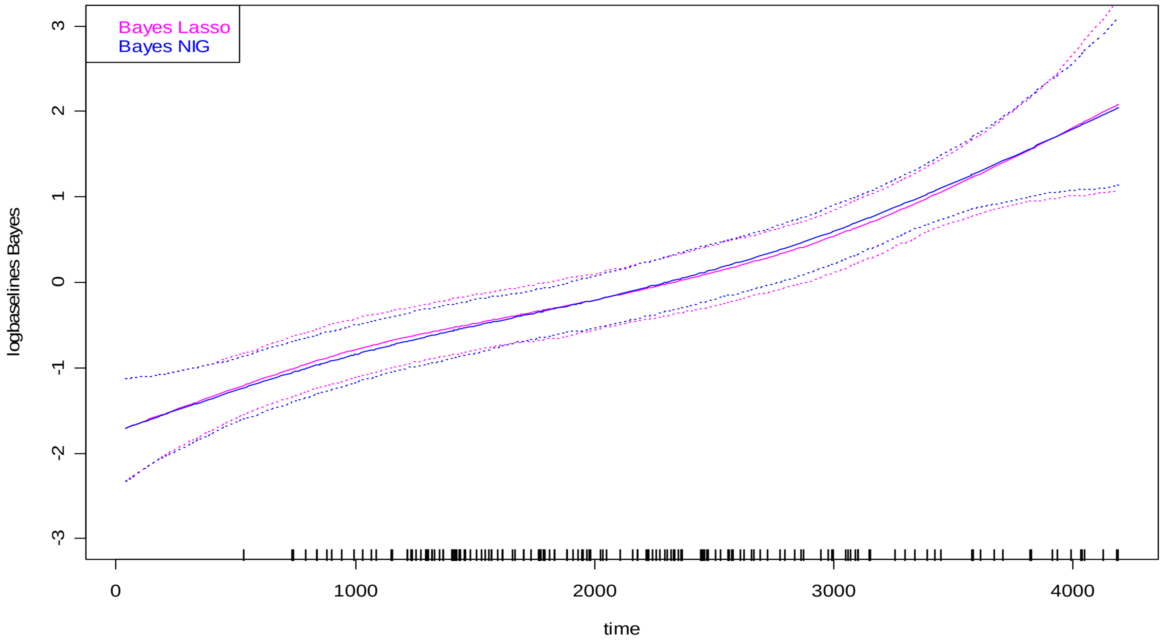

Figure 5.3：Log-baseline hazard (solid lines) posterior mean estimates for the Bayesian NMIG and Bayesian Lasso based on full likelihood with 0.95 pointwise credible bands (dotted lines).

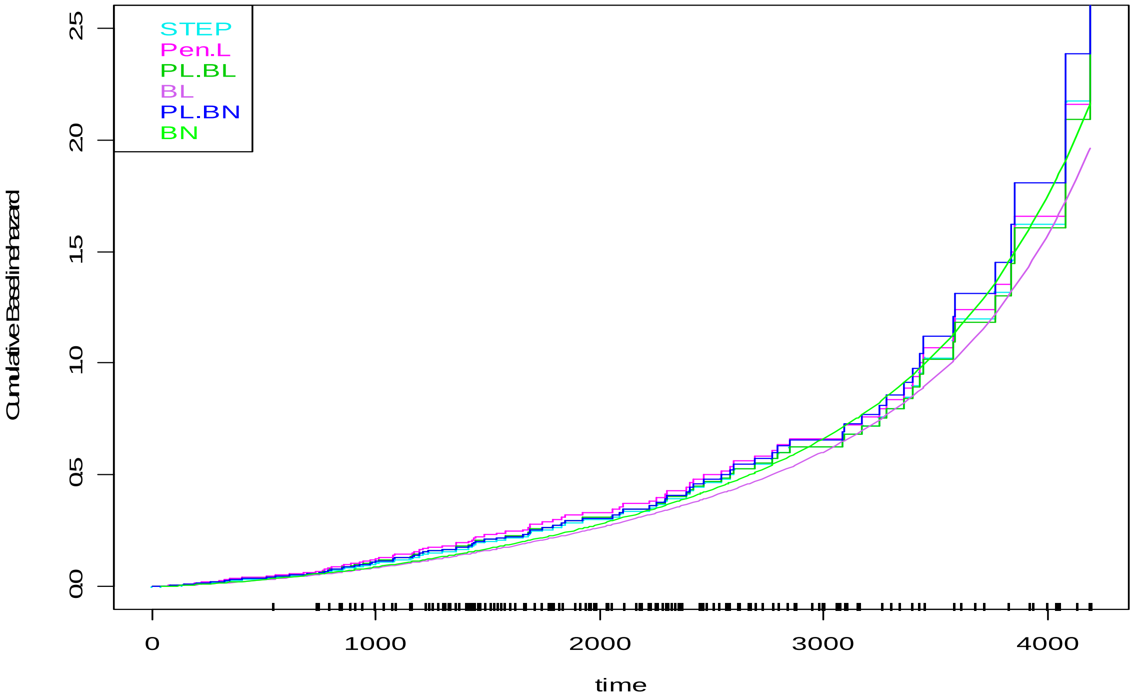

Figure 5.4 Cumulative baseline hazards obtained as Breslow’s estimate for partial likelihood methods and via the trapezoidal rule for full likelihood methods. (SURV: Maximum partial likelihood, PEN.L: Lasso with package penalized, PL.BL: Bayesian Lasso based on partial likelihood, BL: Bayesian Lasso based on full likelihood, PL.BN: Bayesian NMIG based on partial likelihood, BN: Bayesian NMIG based on full likelihood)

In Figure 5.5, the estimated coefficients from the frequentist Lasso are plotted as a function of the standardized constraint parameter $t = \sum |\hat{\beta}_j| / \sum |\hat{\beta}_{j,ML}|$ where $\hat{\beta}_{j,ML}$ denotes the unconstrained partial likelihood estimates and $\hat{\beta}_j$ corresponds to the Lasso estimates with $\sum |\hat{\beta}_j| \le s \in \left[0, \sum |\hat{\beta}_{j,ML}|\right]$. The asterisks on the coefficient paths indicate the values of t for which the coefficients are computed while t increases from 0 to 1. The black dotted vertical line marks the estimated standardized constraint parameter based on n-fold generalized cross validation. We see that the covariates alkphos and sgot enter the final lasso model close to the estimate of the standardized constraint parameter. Figure 5.6 shows the corresponding coefficient paths for the Bayesian Lasso. Figure 5.7 shows the frequencies of the indicator variables as a function of the complexity parameter ω. The vertical line in this figure marks the estimated value $\hat{\omega}$ from the MCMC sample.

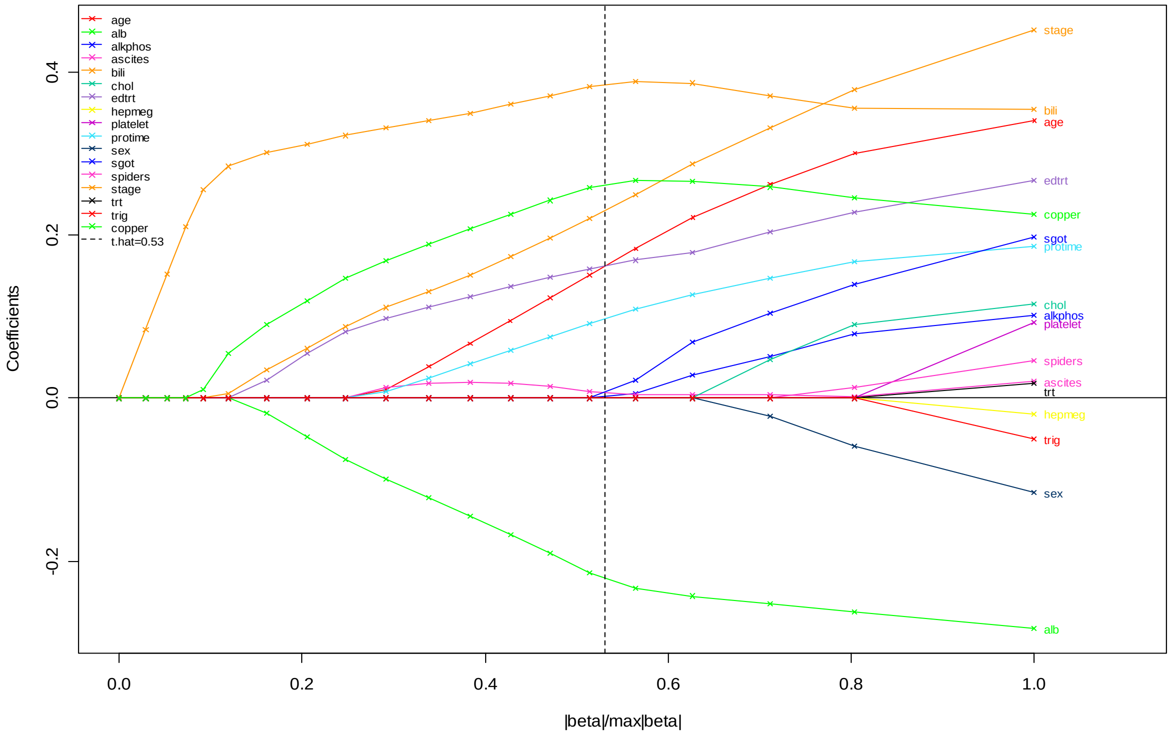

Figure 5.5 Coefficient estimates from the frequentist Lasso as a function of the standardized constraint parameter $t = \sum |\hat{\beta}| / \sum |\hat{\beta}_{ML}|$.

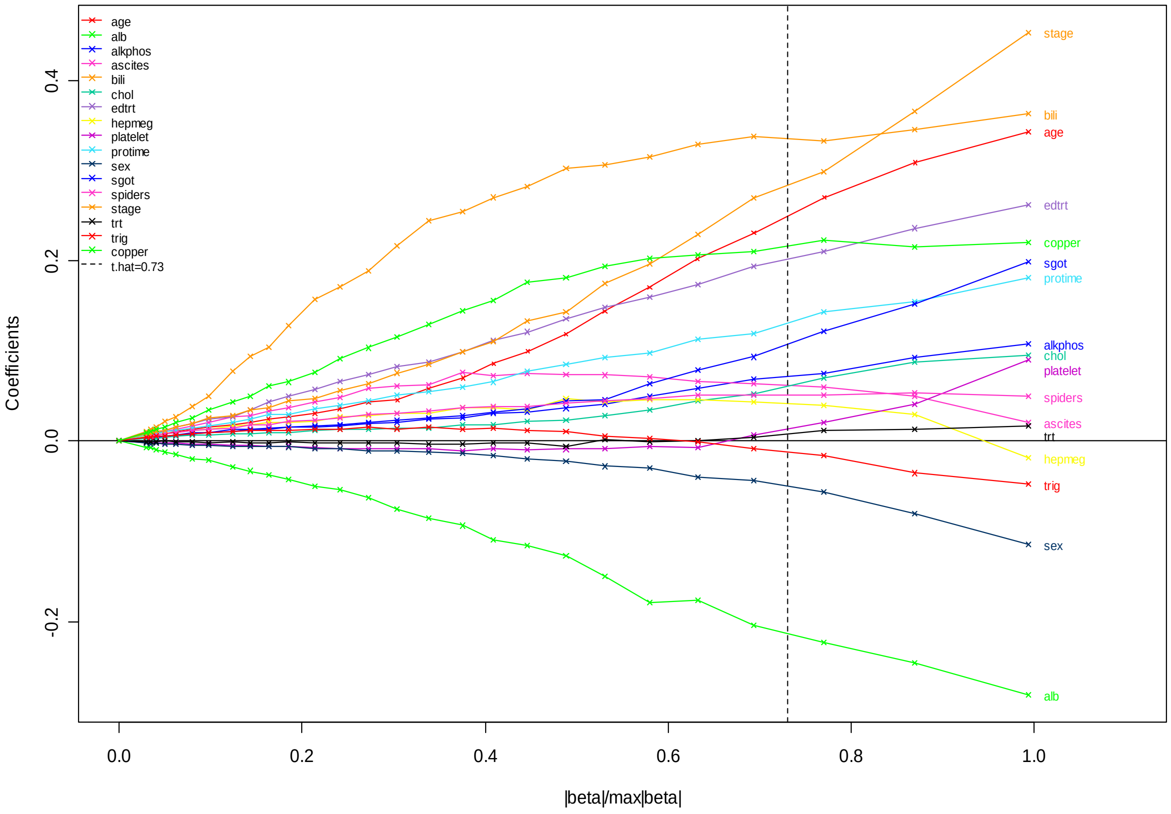

Figure 5.6 Coefficient estimates from the Bayesian Lasso as a function of the standardized constraint parameter $t = \sum|\hat{\beta}|/\sum|\hat{\beta}_{ML}|$.

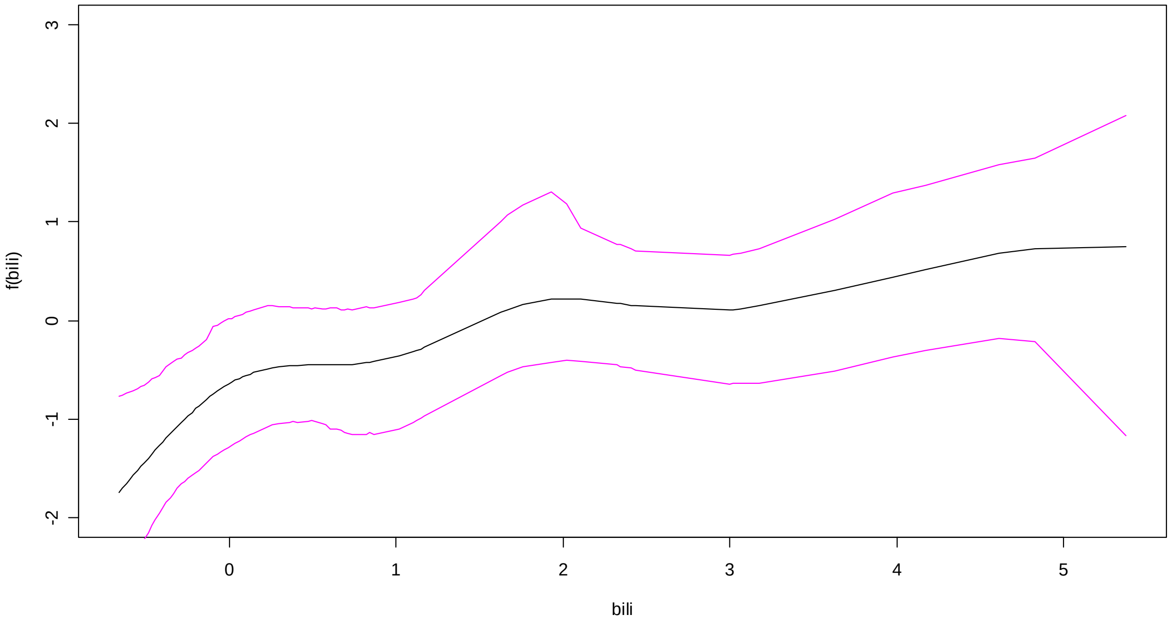

Figure 5.7：Nonlinear effect of the covariate bili modelled with a P-spline of degree 3, 20 knots and a random walk penalty of order 2.

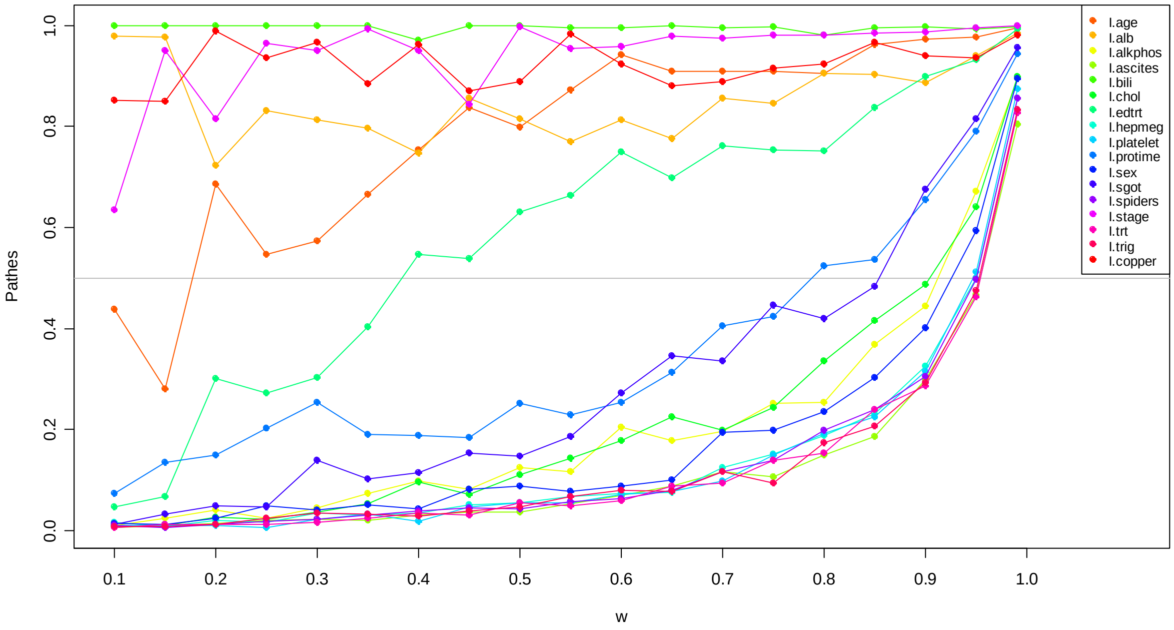

Figure 5.8 : Frequencies of the indicator variables as a function of the complexity parameter ω.

6 Summary and Discussion

We have developed different types of regularization priors for flexible hazard regression models that allow to combine modelling of complex predictor structures with regularization of effects of possibly high-dimensional covariate vectors. Besides classical penalization based regularization priors that mimic the frequentist Ridge or Lasso approach, we considered a normal mixture of inverse gamma distributions as prior that supplements regularization with a natural possibility for variable selection based on latent indicator variables. The basic advantages of the Bayesian formulation of the regularization problem are two-fold: On the one hand, complex models can be built from blocks considered in previous approaches due to the modularity of MCMC simulations. On the other hand, the Bayesian formulation allows for the simultaneous estimation of all parameters involved while allowing for significance and uncertainty statements even about complex functions of these parameters. The Deviance Information Criterion provides a natural means of model comparison, for example between models with and without nonparametric effects. The restriction that posterior mean estimates in regularized regression models do not directly provide the variable selection property known for example from the frequentist Lasso can be overcome by the latent indicator approach in the NMIG prior model. Furthermore, Hans (2008) provides some evidence that posterior mean models without hard shrinkage of coefficients may even be beneficial when considering prediction from regularized regression models, in particular if the sparsity assumption is not fulfilled by the data under consideration. The NMIG prior model also has shown very satisfying properties in our simulation studies and applications. In the future, application in models with more variables than observation will be of obvious interest. In the context of gene expression data, for example, the advantages of the Bayesian approach will be particularly valuable since flexible modelling of clinical covariates can be combined with regularization of high-dimensional microarray features.

In future research, adaptive versions of the proposed regularization priors will be considered. These allow for separate smoothness parameters in addition to the separate variance parameters already included in the scale mixture representation, yielding further flexibility and adaptivity to the scaling of covariates. Hopefully it will therefore be possible to overcome the necessity to standardize covariates up-front, since the priors are allowed to adapt to the varying scaling. The class of regularized regression models for survival times will be broadened by considering accelerated failure time (AFT) models. In the Bayesian formulation, the problem of censoring can be overcome by imputing the unobserved survival times, leading to a regularized linear regression model that is to be fitted in every MCMC iteration. For the special case of log-normal AFT models, the estimation problem then essentially boils down to estimation of Gaussian regression models.

To establish a connection between regularization priors and model choice criteria such as AIC and BIC, Griffin and Brown (2005) compare the log-marginal priors induced by the hierarchical prior formulation with the penalty terms of these model choice criteria. To make the different priors comparable, they are standardized to contain a fixed probability mass within the interval $\left[-2,2\right]$. Griffin and Brown (2005) then establish some relations for Gaussian regression models that could be studied in the context of hazard regression in future research. In addition, investigation of asymptotic properties of the NMIG prior in analogy to the results presented in Ishwaran and Rao (2005) for Gaussian regression models is of obvious interest. In this case, it might be necessary to modify the priors to achieve a non-vanishing impact of the regularization priors even for large sample sizes.

References

Bae, K. and Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics, 20*(18), 3423–3430.

Bender, R., Augustin, T. and Blettner, M. (2005). Simulating survival times for Cox regression models. *Statistics in Medicine, 24,* 1713-1723.

Brezger, A. and Lang, S. (2006). Generalized additive regression based on Bayesian P-Splines, *Computational Statistics & Data Analysis, 50*, 967-991.

Clyde, M. and George, E. (2000): Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society Series B 62,* 681-698.

Cox, D. R. (1972). Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical. Society, Series B 34,* 187-220.

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

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing using B-splines and penalties (with comments and rejoinder). *Statistical Science*, *11*, 89-121.

Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective. *Statistica Sinica*, *14*, 731-761.

Fu, W. J. (1998). Penalized Regressions: The Bridge Versus the Lasso. *Journal of Computational and Graphical Statistics*, *7*(3), 397–416.

Gamerman, D. (1997). Efficient Sampling from the Posterior distribution in Generalized Linear Models. *Statistics and Computing*, *7*, 57-68.

George, E. I. and McCulloch, R. E. (1993). Variable Selection via Gibbs sampling. *Journal of the American Statistical Association. 88*, 881-889.

George, E. I. and McCulloch, R. E. (1997). Approaches For Bayesian Variable Selection. *Statistica Sinica, 7,* 339–373.

Geweke, J. (1996). Variable Selection and Model Comparison in Regression. In Bernardo et al. (eds.): Bayesian Statistics 5, Oxford University Press.

Goeman, J. J. (2007). *An Efficient Algorithm for L1-penalized Estimation.* Leiden, University Medical Center.

Griffin, J. E. and Brown, P. J. (2005). Alternative prior distributions for variable selection with very many more variables than observations. University of Warwick, Deptartment of Statistics, Technical report.

Griffin, J. E. and Brown, P. J. (2007). Bayesian adaptive lassos with non-convex penalization. University of Warwick, Deptartment of Statistics, Technical report.

Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics, 21*(13), 3001–3008.

Hans, C. (2008). *Bayesian Lasso regression*. Technical Report, Department of Statistics, Ohio State University.

Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2006). Geoadditive Survival Models. *Journal of the American Statistical Association*, *101*, 1065-1075.

Ishwaran, H. and Rao, S. J. (2005). Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection. *Journal of the American Statistical Association, (98)*462, 438-455.

Ishwaran, H. and Rao, S. J. (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics, 33,* 730–773.

Jingqin, L. (2006). *Model selection, covariance selection and Bayes classification via shrinkage estimators.* PhD Thesis, Duke University.

Kaderali, L. (2006). *A Hierarchical Bayesian Approach to Regression and its Application to Predicting Survival Times in Cancer.* Dissertation, Universität zu Köln. Köln.

Kaderali, L., Zander, T., Faigle, U., Wolf, J., Schultze, J. L. and Schrader, R. (2006). CASPAR: a hierarchical bayesian approach to predict survival times in cancer from gene expression data. *Bioinformatics, 22*(12), 1495–1502.

Knight, K. and Fu, W. J. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics, 28*(5), 1356–1378.

Lambert, P. and Eilers, P. (2005). Bayesian Survival Models with smooth time-varying coefficients using penalized Poisson regression. *Statistics in Medicine, 24,* 3977–3989.

Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics, 13,* 183-212.

Lee, K. E. (2004). *Bayesian Models for DNA microarray data analysis.* Dissertation, Texas A&M University.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M. and Mallick, B. K. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics, 19,* 90–97.

Lin, D. Y. (2007). On the Breslow estimator. *Lifetime Data Analysis, 13*(4), 471–480.

Lu, W. and Zhang, H. *Variable Selection via Penalized Likelihood with Adaptive Penalty.* Mimeo Series No. 2594, North Carolina State University. North Carolina.

MacLehose, R. F. and Dunson, D. B. (2007). Bayesian semiparametric multiple shrinkage, *to appear.*

Panagiotelis and Smith (2008). Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. *Journal of Econometrics, 143,* 291-316.

Park, M. Y. and Hastie, T. (2006). L1 Regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society, Series B, 69*(4), 659–677.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association, (103)*482, 681-686.

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

Sinha, D., Ibrahim, J. G. and Chen, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika, 90*(3), 629–641.

Smith, M. und Kohn, R. (1996): Nonparametric regression using Bayesian variable selection, *Journal of Econometrics, 75,* 317-343.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B, 64*(4), 583–639.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. Springer, New York.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B, 58*(1), 267–288.

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

van Houwelingen, H. C. (2001). Shrinkage and Penalized Likelihood as Methods to Improve Predictive Accuracy. *Statistica Neerlandica, 55*(1), 17–34.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., Van't Veer, L. J., & Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine, 25,* 3201–3216.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics, 56*(1), 256–262.

Wood, S. N. (2006). *Generalized Additive Models*. Chapman & Hall / CRC.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's Proportional Hazards Model. *Biometrika, 94*(3), 691–703.

Zhang, W., Chaloner, K., Cowles, M. K., Zhang, Y. and Stapleton, J. T. (2008). A Bayesian analysis of doubly censored data using a hierarchical Cox model. *Statistics in Medicine, 27,* 529-542.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B, 67*, 301–320.