Hennerfeind, Brezger, Fahrmeir:

Geoadditive Survival Models: A Supplement

Sonderforschungsbereich 386, Paper 454 (2005)

Online unter: http://epub.ub.uni-muenchen.de/

Geoadditive Survival Models: A Supplement

Andrea Hennerfeind, Andreas Brezger, and Ludwig Fahrmeir*

ABSTRACT

This technical report supplements the paper Geoadditive Survival Models (Hennerfeind, Brezger and Fahrmeir, 2005, Revised for JASA). In particular, we describe the simulation study of this paper in greater detail, present additional results for the application, and provide a complete proof of Theorem 1, Corollary 1, as well as the lemmata and corollaries in the appendix.

1. INTRODUCTION

Hennerfeind et al. (2005) consider Cox–type hazard rate models

$$\lambda_i(t) = exp(\eta_i(t)) \tag{1}$$

with geoadditive predictor

$$\eta_i(t) = g_0(t) + \sum_{j=1}^{p} g_j(t) z_{ij} + \sum_{j=1}^{q} f_j(x_{ij}) + f_{spat}(s_i) + \boldsymbol{v}_i' \boldsymbol{\gamma} + b_{g_i}. \tag{2}$$

Here $g_0(t) = \log\{\lambda_0(t)\}$ is the log–baseline effect, $g_j(t)$ is a time–varying effect of the covariate z_j, $f_j(x_j)$ is the nonlinear effect of a continuous covariate x_j, $f_{spat}(s)$ is the (structured) effect of the spatial covariate s, with $s_i = s$ if subject i is from area s, $s = 1, \ldots, S$, $\boldsymbol{\gamma}$ is the vector of usual linear fixed effects, and b_g is a subject– or group–specific frailty or random effect, with $b_{gi} = b_g$ if individual i is in group g, $g = 1, ..., G$. For $G = n$, we obtain individual–specific frailties, for $G < n$, b_g might be the effect of center g in a multicenter study or the unstructured (uncorrelated random) spatial effect of an area (i.e. $b_g = b_s$), for example. Several other extensions of the model, such as choice of other link functions, inclusion of interactions, random slopes and competing risks, are possible. For identifiability reasons, all unknown functions are centered about zero, and an intercept term is included in the parametric linear term.

Under the usual assumption about noninformative censoring, the likelihood is given by

$$L = \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} \cdot \exp\left(-\int_0^{t_i} \lambda_i(u)du\right) = \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} \cdot S_i(t_i)\,. \tag{3}$$

*Andrea Hennerfeind (email: andreah@stat.uni-muenchen.de), Andreas Brezger (email: andib@stat.uni-muenchen.de) and Ludwig Fahrmeir (email: fahrmeir@stat.uni-muenchen.de) are Statisticians, Ludwig-Maximilians-Universität, Department of Statistics, Ludwigstr. 33, 80539 Munich, Germany. Financial support of the German Science Foundation DFG, Sonderforschungsbereich 386 "Statistische Analyse Diskreter Strukturen" is gratefully acknowledged. The authors thank Dr. H. Hemingway and A. Crook for providing the data for the CABG study, Stefan Lang for helpful discussions and computational assistance with *BayesX*, and Leonhard Held for helpful comments on the CABG data. We also thank the associate editor and the referees for their very useful and constructive comments, helping to substantially improve a first version of this article.

For defining priors and developing posterior analysis the observation model (1) is rewritten in generic matrix notation. Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_i, \ldots, \eta_n)'$ denote the predictor vector, where $\eta_i := \eta_i(t_i)$ is the value of predictor (2) at the observed lifetime $t_i, i = 1, \ldots, n$. Correspondingly, let $\boldsymbol{g}_j = (g_j(t_1), \ldots, g_j(t_n))'$ denote the vector of evaluations of the functions $g_j(t), j = 0, \ldots, p$, $\boldsymbol{f}_j = (f_j(x_{1j}), \ldots, f_j(x_{nj}))'$ the vector of evaluations of the functions $f_j(x_j), j = 1, \ldots, q$, $\boldsymbol{f}_{spat} = (f_{spat}(s_1), \ldots, f_{spat}(s_n))'$ the vector of spatial effects, and $\boldsymbol{b} = (b_{g_1}, \ldots, b_{g_n})'$ the vector of uncorrelated random effects. Furthermore, let $\tilde{\boldsymbol{g}}_j = (g_j(t_1)z_{1j}, \ldots, g_j(t_n)z_{nj})', j = 1, \ldots, p$. Then the vectors $\boldsymbol{g}_0$, $\tilde{\boldsymbol{g}}_j$, $\boldsymbol{f}_j$, $\boldsymbol{f}_{spat}$ and $\boldsymbol{b}$ can always be expressed as the matrix product of an appropriately defined design matrix $\boldsymbol{Z}$, say, and a (possibly high-dimensional) vector $\boldsymbol{\beta}$ of parameters, e.g. $\tilde{\boldsymbol{g}}_j = \boldsymbol{Z}_j\boldsymbol{\beta}_j$, $\boldsymbol{f}_j = \boldsymbol{Z}_j\boldsymbol{\beta}_j$, etc. Then, after reindexing, we can represent the predictor vector $\boldsymbol{\eta}$ in generic notation as

$$\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}_0\boldsymbol{\beta}_0 + \ldots + \boldsymbol{Z}_m\boldsymbol{\beta}_m. \tag{4}$$

For fixed effect parameters $\boldsymbol{\gamma}$ in (4) diffuse priors $p(\boldsymbol{\gamma}) \propto const$ are assumed.

Priors for functions and spatial components are defined by a suitable design matrix $\boldsymbol{Z}_j$, $j = 0, \ldots, m$, and a prior for the parameter vector $\boldsymbol{\beta}_j$. The general form of a prior for $\boldsymbol{\beta}_j$ in (4) is

$$p(\boldsymbol{\beta}_j|\tau_j^2) \propto \tau_j^{-r_j} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right), \tag{5}$$

where $\boldsymbol{K}_j$ is a precision or penalty matrix of $\text{rank}(\boldsymbol{K}_j) = r_j$, shrinking parameters towards zero or penalizing too abrupt jumps between neighboring parameters. For P–splines and MRF priors, $\boldsymbol{K}_j$ will be rank deficient, i.e., $r_j < d_j = \dim(\boldsymbol{\beta}_j)$, and the prior is partially improper.

The variance τ_j^2 acts as an (inverse) smoothing parameter, following inverse Gamma priors $IG(a_j; b_j)$

$$p(\tau_j^2) \propto \frac{1}{(\tau_j^2)^{a_j+1}} \exp\left(-\frac{b_j}{\tau_j^2}\right) \tag{6}$$

to all variances. They are proper for $a_j > 0$, $b_j > 0$, and we use $a_j = b_j = 0.001$ as a standard choice for a weakly informative prior. From our experience results are rather insensitive to the choice of $a_j > 0$ and $b_j > 0$ for moderate to large data sets and the posterior distribution is proper in any case (see Section 4 for a proof). However, since the limiting case, when a_j and b_j are zero, leads to an improper posterior distribution, we present a sensitivity analysis in Section 2 and compare the results to those we obtained with a uniform prior for the standard deviation τ_j, as proposed in Gelman (2004). Note that uniform priors are a special (improper) case of the prior (6) with $a_j = -0.5$, $b_j = 0$, still leading to proper posteriors under regularity assumptions.

The general form (5) of priors covers, among others, Bayesian P–splines for nonlinear effects of a continuous covariate and for time–varying effects, spatial priors in form of MRFs, stationary GRFs and 2d tensor product smoothing splines as well as uncorrelated random intercepts and slopes. For details about these priors as well as for MCMC inference, we refer to Hennerfeind et al. (2005).

2. SIMULATION STUDY

Performance was investigated through simulation studies. In particular we were interested in the following questions: First, how influential is the choice of MRF versus smoother spatial priors, and of an exponential model (P–spline of degree zero) versus a cubic P–spline model for the baseline hazard rate? And second, how sensitive are the results with respect to the hyperparameters for the variance parameters?

Life times T_i, $i = 1, ..., 1236$, were generated according to the hazard model

$$\lambda_i(t) = \lambda_0(t)\exp(f_1(x_i) + f_{spat}(s_i) + \gamma v_i) = \exp(log(3t^2) + sin(x_i) + sin(x_{s_i} \cdot y_{s_i}) - 0.3v_i),$$

with Weibull baseline hazard rate $\lambda_0(t) = 3t^2$, a binary covariate v, with the v_i´s randomly drawn from a Bernoulli $B(1; 0.5)$ distribution, and a continuous covariate x, with the x_i´s randomly drawn from a uniform $U[-3, 3]$ distribution. The spatial covariate s_i denotes one of the $s = 1, \ldots, S = 309$ counties of the former Federal Republic of Germany and x_{s_i} and y_{s_i} are the centered coordinates of the geographic center of county s_i. We simulated four observations per county, resulting in $309 \times 4 = 1236$ observations in total. The censoring was done as follows: We randomly selected a certain proportion of observations ($\approx 17\%$ and $\approx 50\%$, respectively) that were to be censored. Censoring variables C_i for these selected observations were then generated as i.i.d. draws from corresponding uniform $U[0, T_i]$ distributions.

Keeping the predictor fixed, 100 replications $\{T_i^{(r)}, C_i^{(r)}, i = 1, ..., 1236\}$ respectively $\{(t_i^{(r)}, \delta_i^{(r)}), i = 1, ..., 1236\}$, $r = 1, ..., 100$ of censored survival times were generated.

To investigate the first question, the log–baseline hazard $g_0(t)$ was modelled by second order random walk priors, corresponding to a piecewise exponential model, and alternatively as a cubic P–spline with 20 knots. The spatial effect was modelled as a MRF and alternatively as a two–dimensional cubic P–spline with 12×12 knots. Simulations with GRF priors are not feasible due to much higher computation times, but the general message will be the same. A cubic P–spline prior with 20 knots was chosen for $f_1(x) = sin(x)$ in each case. Hyperparameters of inverse Gamma priors for variance components were set to $a = 0.001$, $b = 0.001$, the standard choice.

For each replication $r = 1, ..., 100$, we computed the mean square errors

$$MSE_r(g_0) = \frac{1}{1236}\sum_{i=1}^{1236}(\widehat{g}_0^{(r)}(t_i^{(r)}) - g_0(t_i^{(r)}))^2,$$

for the log–baseline hazard $g_0(t)$,

$$MSE_r(f_1) = \frac{1}{1236}\sum_{i=1}^{1236}(\widehat{f}_1^{(r)}(x_i) - f_1(x_i))^2$$

for $f_1(x) = sin(x)$, and

$$MSE_r(f_{spat}) = \frac{1}{1236}\sum_{i=1}^{1236}(\widehat{f}_{spat}^{(r)}(s_i) - f_{spat}(s_i))^2$$

Table 1: Summary of MSEs

	p.e.m. 17%		P–spline–m. 17%	
MSE–type	MRF	2d P–spline	MRF	2d P–spline
$meanMSE(g_0)$	0.200	0.197	0.164	0.163
$minMSE(g_0)$	0.056	0.054	0.041	0.036
$maxMSE(g_0)$	0.433	0.441	0.385	0.389
$meanMSE(f_1)$	0.0076	0.0063	0.0080	0.0065
$minMSE(f_1)$	0.0014	0.0008	0.0015	0.0009
$maxMSE(f_1)$	0.0241	0.0197	0.0278	0.0198
$meanMSE(f_{spat})$	0.042	0.022	0.043	0.022
$minMSE(f_{spat})$	0.028	0.010	0.028	0.010
$maxMSE(f_{spat})$	0.065	0.042	0.066	0.045
$meanMSE(\gamma)$	0.0059	0.0051	0.0064	0.0051
$minMSE(\gamma)$	≈ 0	≈ 0	≈ 0	≈ 0
$maxMSE(\gamma)$	0.0365	0.0255	0.0379	0.0247
	p.e.m. 50%		P–spline–m. 50%	
MSE–type	MRF	2d P–spline	MRF	2d P–spline
$meanMSE(g_0)$	0.534	0.527	0.456	0.446
$minMSE(g_0)$	0.314	0.312	0.237	0.217
$maxMSE(g_0)$	0.923	0.916	0.810	0.844
$meanMSE(f_1)$	0.0168	0.0132	0.0175	0.0140
$minMSE(f_1)$	0.0006	0.0011	0.0005	0.0009
$maxMSE(f_1)$	0.0591	0.0396	0.0605	0.0429
$meanMSE(f_{spat})$	0.055	0.031	0.056	0.032
$minMSE(f_{spat})$	0.032	0.013	0.032	0.013
$maxMSE(f_{spat})$	0.099	0.064	0.107	0.066
$meanMSE(\gamma)$	0.0104	0.0086	0.0110	0.0087
$minMSE(\gamma)$	≈ 0	≈ 0	≈ 0	≈ 0
$maxMSE(\gamma)$	0.0507	0.0489	0.0497	0.0509

Table 2: Summary of MSEs

prior	IG,a=b=0.001	IG,a=b=0.0001	IG,a=b=1e-05	IG,a=b=1e-08	uniform
$meanMSE(g_0)$	0.164	0.164	0.165	0.166	0.157
$minMSE(g_0)$	0.041	0.039	0.038	0.038	0.032
$maxMSE(g_0)$	0.385	0.391	0.383	0.385	0.363
$meanMSE(f_1)$	0.0080	0.0079	0.0079	0.0079	0.0085
$minMSE(f_1)$	0.0015	0.0015	0.0015	0.0016	0.0019
$maxMSE(f_1)$	0.0278	0.0269	0.0280	0.0287	0.0296
$meanMSE(f_{spat})$	0.043	0.043	0.043	0.043	0.043
$minMSE(f_{spat})$	0.028	0.027	0.027	0.028	0.028
$maxMSE(f_{spat})$	0.066	0.065	0.067	0.066	0.066
$meanMSE(\gamma)$	0.0064	0.0064	0.0063	0.0063	0.0064
$minMSE(\gamma)$	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
$maxMSE(\gamma)$	0.0379	0.0398	0.0380	0.0392	0.0390

for the spatial effect $f_{spat}(s) = sin(x_c \cdot y_c)$, where $\widehat{g}_0^{(r)}$ and $\widehat{f}_k^{(r)}$, $k = 1, spat$, are posterior mean estimates for simulation run r. The $MSE(\gamma)$ was computed in the usual way.

Table 1 summarizes the results, displaying $meanMSE = \frac{1}{100}(\sum_{r=1}^{100} MSE_r)$ as well as $min_r MSE_r$ and $max_r MSE_r$ in each cell. As was to be expected, the P–spline model has smaller MSEs for g_0 when compared to the piecewise exponential model. Interestingly, the MSEs for $\gamma = -0.3$, $f_1(x)$ and $f_{spat}(s)$ are more or less unaffected by the choice of the smoothness prior for the log–baseline $g_0(t)$. Estimated functions of replication r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$, for $g_0(t)$, $f_1(x)$ and $f_{spat}(s)$ are displayed in Figures 1-3 (for the censoring level of 17%). Regarding the two different levels of censoring Tables 1 shows that the estimation of the log–baseline effect is the effect that is strongest influenced by the level of censoring. While increasing the censoring level from 17% to 50% leads to an approximately 2.75 times larger MSE for $g_0(t)$ the MSE for $f_{spat}(s)$ is only increased by a factor of ca. 1.35.

In order to analyze the behavior of the Markov chains when a and b approach zero (and the prior for the hyperparameters thus approaches the $IG(0;0)$ distribution, that leads to an improper posterior), we exemplary single out the P–spline model with MRF-prior and a censoring level of 17% and alternatively set $a = b = 0.0001$, $a = b = 0.00001$ and $a = b = 0.00000001$. We additionally run the simulation study with uniform priors (i.e. $a = -0.5$, $b = 0$) on the standard deviations τ_0, τ_1 and τ_{spat} that act as smoothing parameters for the log-baseline, the nonlinear effect of x and the spatial effect, respectively. Selected sampling paths of run $r = 1$ are exemplary shown in Figure 4. We did not face problems with mixing or convergence of Markov chains with any of these prior distributions. An exception are the first one or two parameters of the baseline effect, i.e. $\beta_{0,1}$ and $\beta_{0,2}$, corresponding to the effect of small times t, where the mixing properties are not always optimal. This can be explained by the very steep increase of the 'true' log–baseline, reaching to minus infinity as t approaches zero whereas it is quite flat elsewhere. In this situation a global variance might not be an ideal choice. Another point may be the usage of conditional prior proposals that usually lead to poorer mixing properties than IWLS–proposals do. Figure 5 displays kernel density estimators of the posterior mean of the variance parameters based on $\widehat{\tau_j^2}^{(r)}, r = 1, \ldots, 100$ for $j = 0, 1, spat$. Obviously the different choices of the hyperparameters a and b of the inverse Gamma prior do not seem to have much effect, whereas the uniform prior on the standard deviations tends to result in larger estimates for the variance parameters and thus in less smooth effects. The posterior distribution of the variance parameter of the spatial effect is less sensitive to the different choices of priors, as the full conditional is dominated by the values of r_j =rank($\boldsymbol{K}_j$) and $\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j$ at this. Table 2 summarizes the MSEs, that are computed and displayed as before. While the MSEs are quite unaffected by the choice of the hyperparameters a and b of the inverse Gamma prior, the uniform prior results in a slightly smaller MSE for $g_0(t)$, but a slightly bigger MSE for $f_1(x)$. Altogether we come to the conclusion that (at least with this model) it does not seem to be crucial, which one of these weakly informative priors is assumed for the variance parameters.

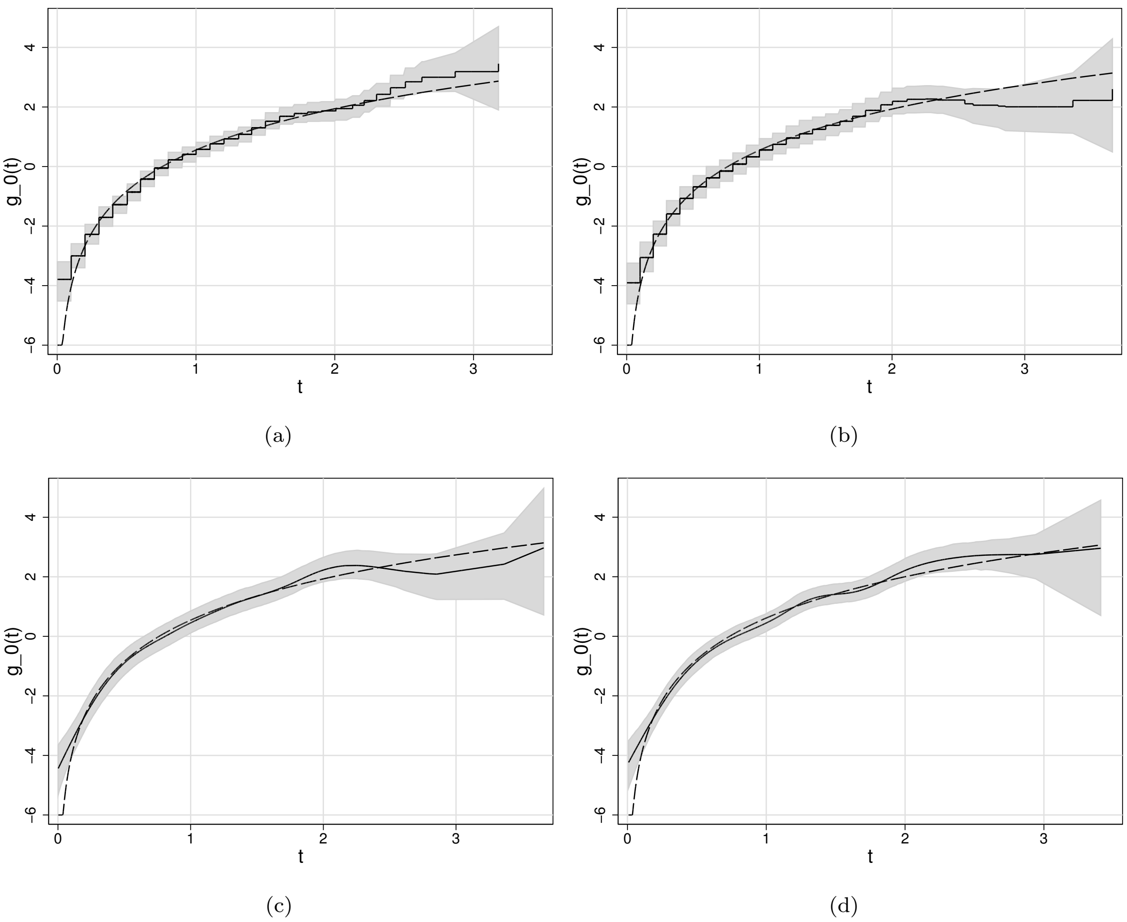


Figure 1: (log–)Baseline effects for the various model specifications; displayed are posterior mean estimates and 95% credible intervals of run r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$ (solid line and grey shaded area), and the true (log–)baseline effect (dashed line). a) p.e.m., MRF, r=11, MSE=0.183 b) p.e.m., 2d P–spline, r=51, MSE=0.181 c) P–spline model, MRF, r=51, MSE=0.148 d) P–spline model, 2d P–spline, r=7, MSE=0.145

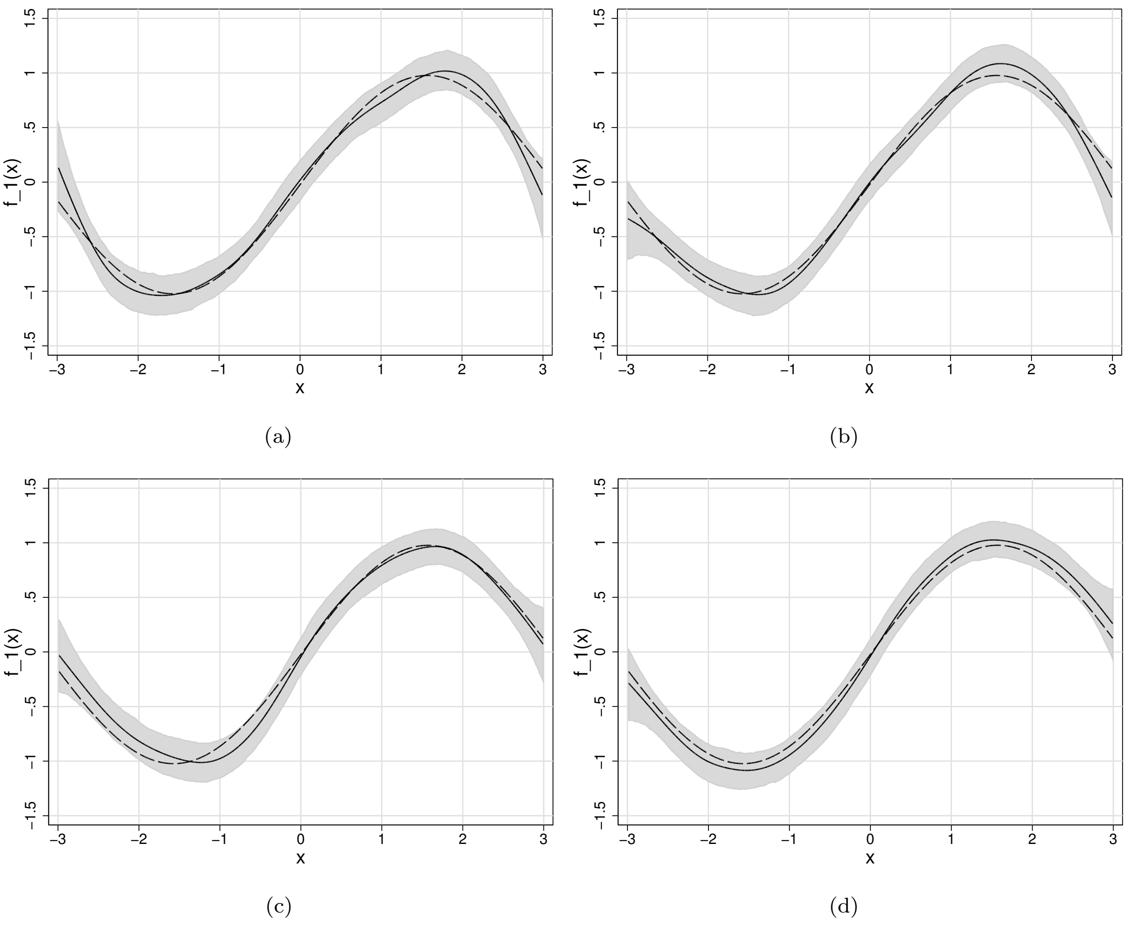


Figure 2: Nonparametric effects for the various model specifications; displayed are posterior mean estimates and 95% credible intervals of run r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$ (solid line and grey shaded area), and the true function (dashed line). a) p.e.m., MRF, r=53, MSE=0.0064 b) p.e.m., 2d P–spline, r=36, MSE=0.0053 c) P–spline model, MRF, r=67, MSE=0.0068 d) P–spline model, 2d P–spline, r=19, MSE=0.0056

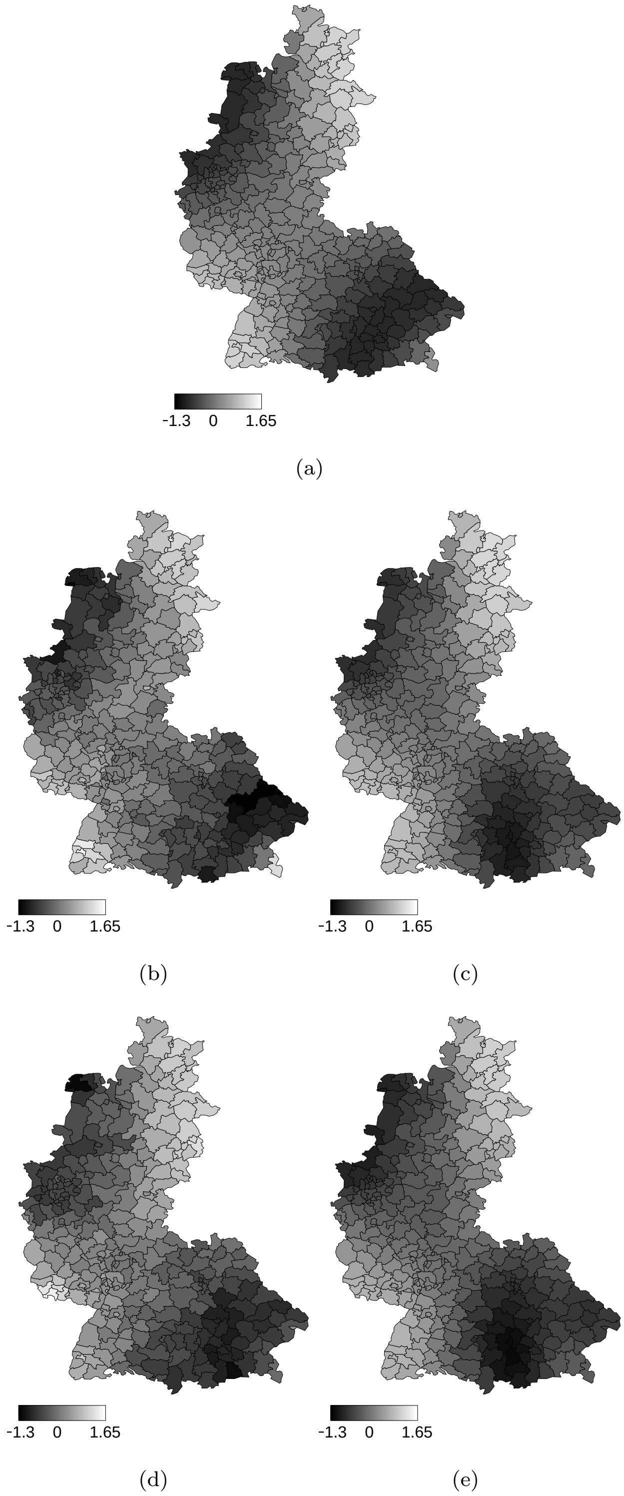


Figure 3: Spatial effects for the various model specifications; displayed are posterior mean estimates of run r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$ a) true function b) p.e.m., MRF, r=41, MSE=0.041 c) p.e.m., 2d P–spline, r=13, MSE=0.021 d) P–spline model, MRF, r=12, MSE=0.042 e) P–spline model, 2d P–spline, r=13, MSE=0.021

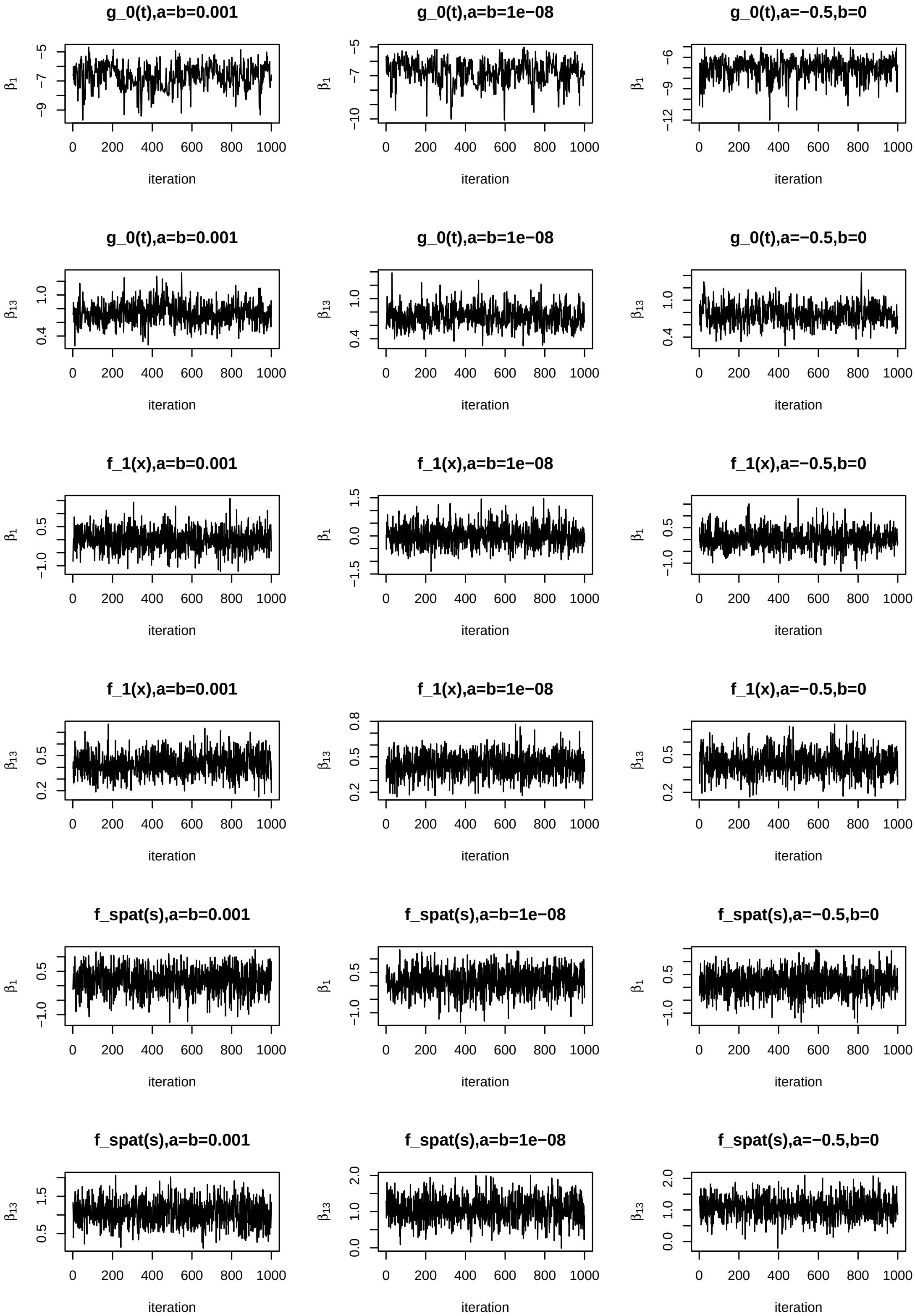


Figure 4: Selected sampling paths of run $r = 1$ for parameters $\beta_{j,1}$ and $\beta_{j,13}$, $j = 0, 1, spat$ and different choices for the parameters a and b of the $IG(a; b)$ hyperpriors.

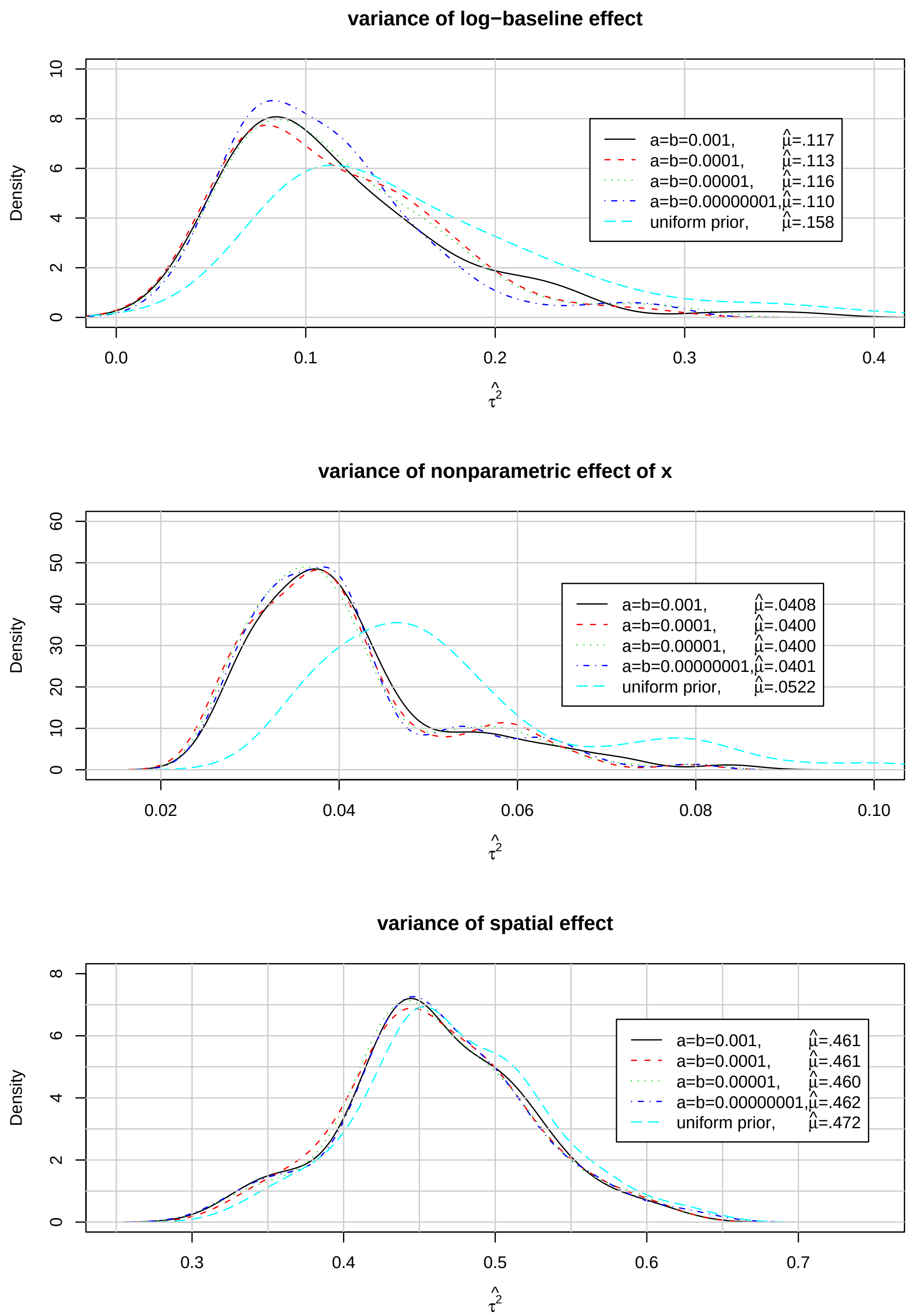


Figure 5: Kernel density estimates based on $\widehat{\tau_j^2}^{(r)}$, $r = 1, \ldots, 100$ for $j = 0, 1$ and *spat*, respectively. $\hat{\mu}$ denotes the mean estimated smoothing parameter.

3. APPLICATION: WAITING TIMES TO CABG

In Hennerfeind et al. (2005) we illustrate our methods by an application to data from a study in London and Essex that aims to analyze the effects of area of residence and further individual specific covariates on waiting times to coronary artery bypass graft (CABG). The data comprise observations for 3015 patients with definite coronary artery disease who were referred to one cardiothoracic unit from five contiguous health authorities. Waiting times from angiography to CABG are given in days. Covariates are, among others, sex, age (in years), number of diseased vessels (1, 2, 3), and the area of residence (one of 488 electoral wards). We analyzed and compared a hierarchy of models, with model comparison based on the deviance information criterion (DIC), but we concentrated on $IG(0.001; 0.001)$ priors for the variances. In this supplement we exemplary present some additional results of model 8 in Hennerfeind et al. (2005) that were obtained with other choices of $IG(a; b)$ priors. Model 8 corresponds to a model with hazard rate

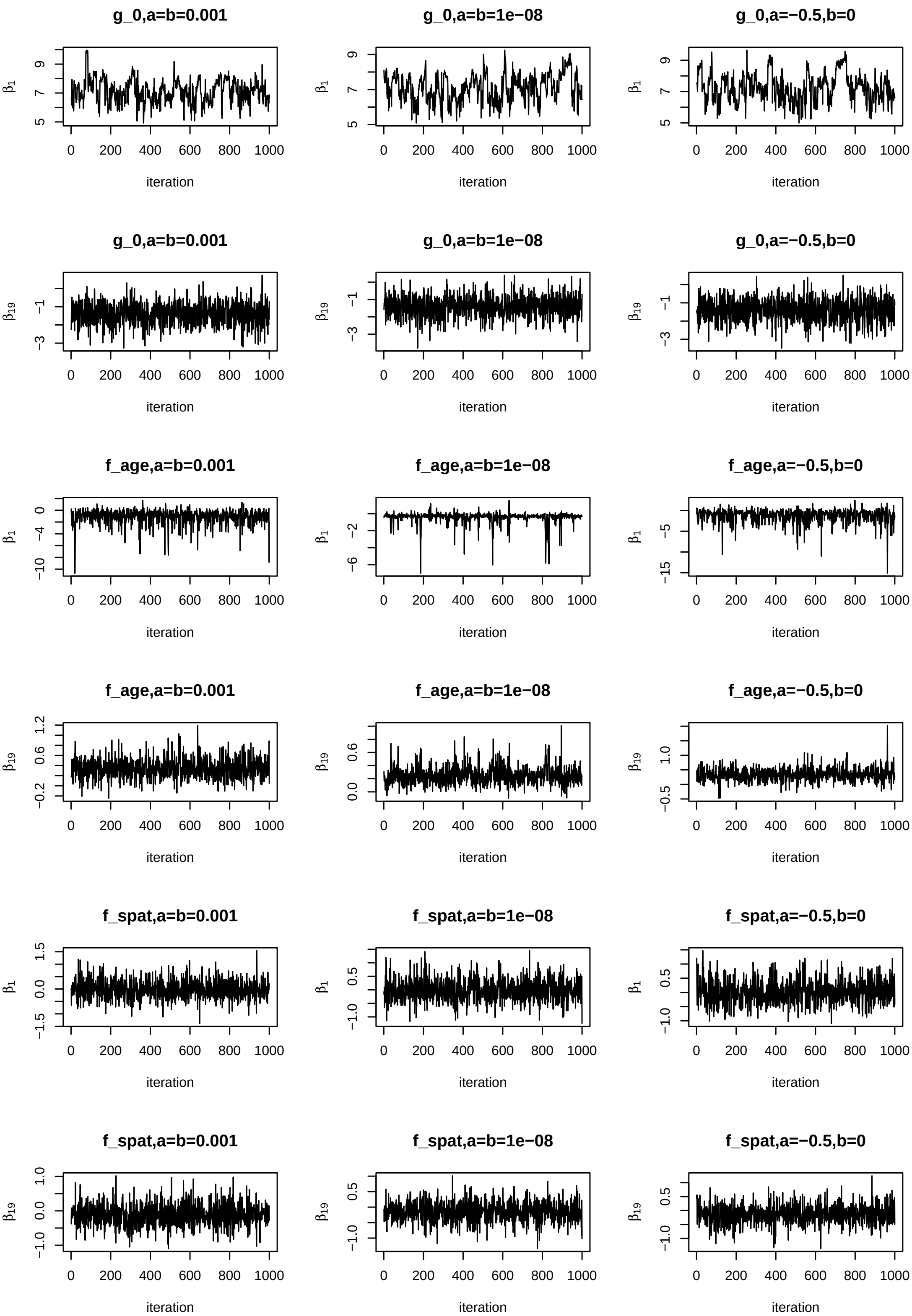
$$\lambda(t) = \exp(g_0(t) + f_{age}(age) + f_{spat}(ward) + \gamma_1 sex + \gamma_2 dv2 + \gamma_3 dv3),$$

where $g_0(t)$ is the log–baseline rate, $f_{age}(age)$ is the nonlinear effect of age and $f_{spat}(ward)$ is the structured spatial effect. The remaining covariates are dummy–coded: $sex = 1$ for female, and $sex = 0$ for male, $dv2 = 1$ if the number of diseased vessels equals 2, $dv2 = 0$ else, and $dv3 = 1$ if the number of diseased vessels equals 3, $dv3 = 0$ else.

For $g_0(t)$ and f_{age} we assumed cubic P–spline priors with 20 knots and the spatial effect $f_{spat}(ward)$ is modelled through a MRF prior. Inverse Gamma priors $IG(a; b)$ were assumed for the variances. In addtion to our standard choice $a = b = 0.001$ we set $a = b = 1e - 08$ and $a = -0.5$, $b = 0$ (i.e. uniform prior on the standard deviation).

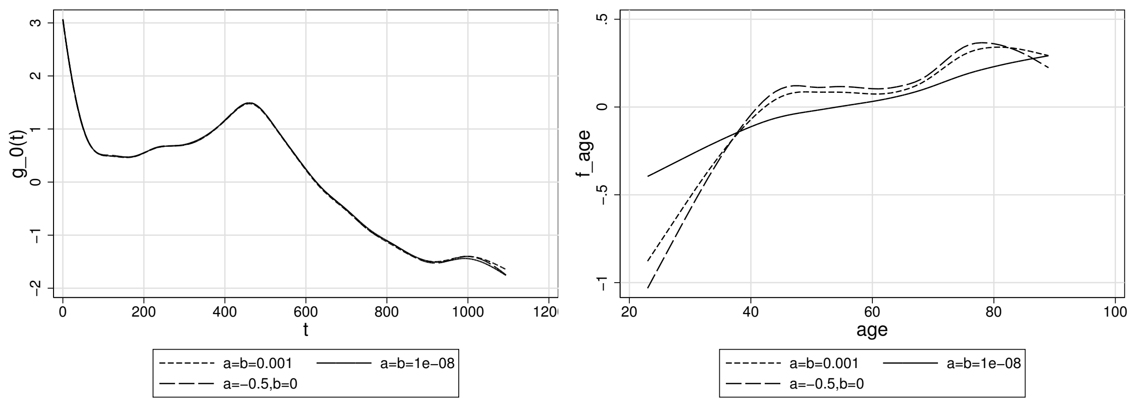
Figure 6 exemplary shows sampling paths of the first and 19th parameter of each vector $\boldsymbol{\beta}_j$, $j = 0, age, spat$ corresponding to the log–baseline effect, the effect of age and the spatial effect, respectively. Independently of the choice of the prior for the hyperparameters the mixing is not optimal for the first parameters of the parameter–vector $\boldsymbol{\beta}_0$ corresponding to the log–baseline effect. In accordance with our simulation study this might be due to the usage of conditional prior proposals and the assumption of a global variance, since the effect is steeply dropping in the first 100 days, but comparatively flat elsewhere. Apart from that we did not face problems with mixing or convergence in the case of $IG(0.001; 0.001)$ and $IG(-0.5; 0)$ priors. However, in the case of an $IG(1e - 08; 1e - 08)$ prior mixing properties are poor for the first parameters of the effect of age, where we have sparse data since there is only a very small number of young patients that suffer from coronary artery diseases. As shown in Figure 7 a) the estimated log–baseline effects $g_0(t)$ are not influenced by the choice of the hyperprior. The same applies to the fixed effects as well as the spatial effect. Figure 7 b) however reveals a much smoother effect with the $IG(1e - 08; 1e - 08)$ prior compared to the effects the other two choices for the hyperpriors yield. But since credible intervals are quite large, each estimated effect is within the 95% credible interval of each other estimated effect of age.

We conclude that the results are in general quite insensitive regarding the choice of non–informative hyper-



(a)

Figure 6: Selected sampling paths for parameters $\beta_{j,1}$ and $\beta_{j,19}$, $j = 0, age, spat$ and different choices for the parameters a and b of the $IG(a; b)$ hyperpriors.



(a) (b)

Figure 7: Estimated log–baseline effects $g_0(t)$ and effects of age f_{age} with different specifications of $IG(a;b)$ hyperpriors.

priors. However, in situations where data are sparse $IG(a;b)$ priors with a and b close to zero might lead to poor mixing and are therefore not recommended.

4. PROOFS OF PROPRIETY RESULTS

The proofs of Theorem 1 and Corollary 1 on propriety of posteriors for geoadditive survival models in the Appendix of Hennerfeind et al. (2005) are based on lemmata extending propriety results for (generalized) mixed models in Sun et al. (1999).

We first consider Gaussian linear mixed models

$$\boldsymbol{y} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}_1\boldsymbol{\beta}_1 + ... + \boldsymbol{Z}_m\boldsymbol{\beta}_m + \boldsymbol{\varepsilon} \tag{7}$$

for observations $\boldsymbol{y} = (y_1, ..., y_n)'$, with a Gaussian error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n) \sim N(\boldsymbol{0}, \tau_0^2\boldsymbol{I})$. For identifiability reasons, the predictor must not contain individual–specific uncorrelated random effects in addition to $\boldsymbol{\varepsilon}$. The prior assumptions for the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_j$, $j = 1, ..., m$, are the same as in Section 2, i.e., a flat prior

$$p(\boldsymbol{\gamma}) \equiv 1 \tag{8}$$

for the vector $\boldsymbol{\gamma}$ of 'fixed' effects, and prior (5) for $\boldsymbol{\beta}_j$. Priors for hyperparameters $\boldsymbol{\tau}^2 = (\tau_0^2, ..., \tau_m^2)'$ are $p(\boldsymbol{\tau}^2) = \prod_{j=0}^m p(\tau_j^2)$. An important special case are inverse Gamma priors (6), which are proper for $a_j > 0$, $b_j > 0$.

Defining $\boldsymbol{Z} = (\boldsymbol{Z}_1, ..., \boldsymbol{Z}_m)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_m')'$, the model (7) is

$$\boldsymbol{y} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Further, with $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{Z})$, let $(\hat{\gamma}', \hat{\boldsymbol{\beta}}')' = (\boldsymbol{X'X})^{-}\boldsymbol{X'y}$ be the least squares estimator, and

$$SSE = \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-}\boldsymbol{X'})\boldsymbol{y}$$

be the sum of squared errors, which is invariant for any choice of the generalized inverse $\boldsymbol{(X'X)^{-}}$.

Lemma A1

Consider the Gaussian mixed model defined by (7), (8) and (5), and assume that the following conditions hold:

(i) rank($\boldsymbol{V}$)=p, rank($\boldsymbol{Z'RZ}+\boldsymbol{K}$)=$d$
where $p=\dim(\boldsymbol{\gamma})$, $d=d_1+...+d_m=\dim(\boldsymbol{\beta})$, $\boldsymbol{K}=\text{diag}(\boldsymbol{K_1},...,\boldsymbol{K_m})$, $\boldsymbol{R=I-V(V'V)^{-1}V'}$.

(ii) the priors $p(\tau_j^2)$, $j=1,...,m$ are proper, and

$$\int p(\tau_0^2)\tau_0^{-(n-p-(d-r))}\exp\left(-\frac{SSE}{2\tau_0^2}\right)d\tau_0^2<\infty,$$

where $r=r_1+...+r_m$.

Then the posterior distribution $p(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\tau^2}\mid\boldsymbol{y})$ is proper.

Corollary A1

For a linear mixed model (7) with prior (8) and (6), the posterior $p(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\tau^2}\mid\boldsymbol{y})$ is proper if condition (i) of Lemma A1 and

$$a_j>0,\ b_j>0,\quad j=1,...,m,$$

$$n-p-(d-r)+2a_0>0,\quad SSE+2b_0>0$$

hold.

Remark: Condition (i) of Lemma A1 is equivalent to

$$\text{rank}\begin{pmatrix}\boldsymbol{V'V} & \boldsymbol{V'Z}\\ \boldsymbol{Z'V} & \boldsymbol{Z'Z+K}\end{pmatrix}=p+d.$$

Proof of Lemma A1 and Corollary A1

The proof extends arguments in Sun, Tsutakawa and Speckman (1999), see also Speckman and Sun (2003), using a theorem on eigenvalues in Magnus and Neudecker (1991). From the model assumptions we have

$$p(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\tau^2}\mid\boldsymbol{y})\propto\tau_0^{-n}\tau_1^{-r_1}\cdot...\cdot\tau_m^{-r_m}\cdot\exp\left\{-\frac{\boldsymbol{(y-V\gamma-Z\beta)'(y-V\gamma-Z\beta)}}{2\tau_0^2}-\sum_{j=1}^{m}\frac{\boldsymbol{\beta_j'K_j\beta_j}}{2\tau_j^2}\right\}p(\boldsymbol{\tau^2})$$

Following Sun et al. (1999), we rewrite

$$\boldsymbol{(y-V\gamma-Z\beta)'(y-V\gamma-Z\beta)}=SSE+(\gamma-\hat{\gamma}-\boldsymbol{c_1})'\boldsymbol{V'V}(\gamma-\hat{\gamma}-\boldsymbol{c_1})+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{Z'RZ}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}),$$

where $\boldsymbol{c_1}=\boldsymbol{(V'V)^{-1}V'Z}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$.

Integrating the right hand side with respect to $\boldsymbol{\gamma}$, we get

$$\int p(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\tau^2}\mid\boldsymbol{y})d\boldsymbol{\gamma}\propto\frac{(2\pi)^{p/2}\mid\boldsymbol{V'V}\mid^{-1/2}}{\tau_0^{n-p}\prod_{j=1}^m\tau_j^{r_j}}\cdot\exp\left(-\frac{SSE}{2\tau_0^2}-\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\boldsymbol{Z'RZ}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{2\tau_0^2}-\frac{1}{2}\boldsymbol{\beta'K_{\tau^2}\beta}\right)p(\boldsymbol{\tau^2}),$$

where $\boldsymbol{K}_{\tau^2}=\text{diag}\left(\boldsymbol{K_1}/\tau_1^2,...,\boldsymbol{K_m}/\tau_m^2\right)$.

Define $\boldsymbol{R_1}= \tau_0^{-2}\boldsymbol{Z'RZ}+\boldsymbol{K_{\tau^2}}$. Then for any $\tau_j^2>0$, $j=0,...,m$, $\boldsymbol{R_1^{-1}}$ exists by assumption (i) of Lemma A1. Set

$$\begin{aligned} \boldsymbol{c_2} &= \tau_0^{-2}\boldsymbol{R_1^{-1}Z'RZ}\hat{\boldsymbol{\beta}} \\ \boldsymbol{R_2} &= \boldsymbol{Z'RZ}-\tau_0^{-2}\boldsymbol{Z'RZR_1^{-1}Z'RZ}. \end{aligned}$$

Then

$$\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\boldsymbol{Z'RZ}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{\tau_0^2}+\boldsymbol{\beta' K_{\tau^2}\beta}=(\boldsymbol{\beta}-\boldsymbol{c_2})'\boldsymbol{R_1}(\boldsymbol{\beta}-\boldsymbol{c_2})+\frac{\hat{\boldsymbol{\beta}}'\boldsymbol{R_2}\hat{\boldsymbol{\beta}}}{\tau_0^2}.$$

Integrating out $\boldsymbol{\beta}$, we get

$$\int p(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\tau^2}\mid \boldsymbol{y})d\boldsymbol{\gamma}d\boldsymbol{\beta}\propto \frac{(2\pi)^{(p+r)/2}\mid \boldsymbol{V'V}\mid^{-\frac{1}{2}}\mid \boldsymbol{R_1}\mid^{-\frac{1}{2}}}{\tau_0^{n-p}\prod_{j=1}^m \tau_j^{r_j}}\cdot \exp\left\{-\frac{SSE+\hat{\boldsymbol{\beta}}'\boldsymbol{R_2}\hat{\boldsymbol{\beta}}}{2\tau_0^2}\right\}p(\boldsymbol{\tau^2}). \tag{9}$$

Since $\boldsymbol{R_2}$ is nonnegative definite, the second factor is bounded by $\exp\left\{-SSE/(2\tau_0^2)\right\}$.
For an upper bound of the first factor, we first derive a lower bound for $\mid \boldsymbol{R_1}\mid$, applying Theorem 9 in Magnus and Neudecker (1991, ch. 11, p. 208) to the eigenvalues of

$$\boldsymbol{R_1}=\tau_0^{-2}\boldsymbol{Z'RZ}+\boldsymbol{K_{\tau^2}}.$$

Note that the $d-r$ smallest eigenvalues of $\boldsymbol{K}$ and $\boldsymbol{K_{\tau^2}}$ are zero, while the eigenvalues $\lambda_l(\boldsymbol{K_{\tau^2}})$, $l=d-r+1,...,r$, are positive. Application of the theorem to the positive eigenvalues of $\boldsymbol{R_1}$ gives

$$\lambda_l(\tau_0^{-2}\boldsymbol{Z'RZ}+\boldsymbol{K_{\tau^2}})\geq \lambda_l(\boldsymbol{K_{\tau^2}})=\lambda(\boldsymbol{K_j})\tau_j^{-2}\geq \lambda_j\tau_j^{-2},$$

where $\lambda(\boldsymbol{K_j})$ is a positive eigenvalue of one of the precision matrices $\boldsymbol{K_j}$ and $\lambda_j>0$ is the smallest positive eigenvalue of $\boldsymbol{K_j}$.
Application of the theorem to the eigenvalues $\lambda_l(\boldsymbol{K_{\tau^2}})=0$, $l=1,...,d-r$, of $\boldsymbol{K_{\tau^2}}$ gives

$$\lambda_l(\tau_0^{-2}\boldsymbol{Z'RZ}+\boldsymbol{K_{\tau^2}})\geq \lambda_l(\tau_0^{-2}\boldsymbol{Z'RZ})\geq \tau_0^{-2}\lambda_0,$$

where $\lambda_0>0$ is the smallest eigenvalue of $\boldsymbol{Z'RZ}$.
Taken together, we get

$$\mid \boldsymbol{R_1}\mid=\prod_l \lambda_l(\boldsymbol{R_1})\geq \tau_0^{-2(d-r)}\prod_{j=1}^m \tau_j^{-2r_j}\cdot L$$

where $L=\lambda^{d-r}\prod_{j=1}^m \lambda_j^{r_j}>0$, and

$$\mid \boldsymbol{R_1}\mid^{-1/2}\leq \frac{1}{L^{1/2}}\tau_0^{d-r}\prod_{j=1}^m \tau_j^{r_j}.$$

Inserting in (9), we obtain

$$\int p(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\delta}\mid \boldsymbol{y})d\boldsymbol{\gamma}d\boldsymbol{\beta}\leq C\frac{1}{\tau_0^{n-p-(d-r)}}\cdot \exp\left\{-\frac{SSE}{2\tau_0^2}\right\}\prod_{j=0}^m p(\tau_j^2).$$

Thus, if condition (ii) in Lemma A1 holds, the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\tau^2} \mid \boldsymbol{y})$ is proper.

Corollary A1 follows immediately, because then

$$\frac{1}{\tau_0^{n-p-(d-r)}} \exp\left\{-\frac{SSE}{2\tau_0^2}\right\} \frac{1}{(\tau_0^2)^{a_0+1}} \exp\left\{-\frac{b_0}{\tau_0^2}\right\} = \frac{1}{\tau_0^{n-p-(d-r)+2(a_0+1)}} \exp\left\{-\frac{SSE/2+b_0}{\tau_0^2}\right\}.$$

We recognize a proper inverse Gamma density for $(n-p-(d-r))/2+a_0>0$ and $SSE/2+b_0>0$.

Propriety of the posterior for generalized (geo–) additive models

The following Lemma A2 gives sufficient conditions for the propriety of the posterior in generalized linear and additive mixed models. The lemma and its proof rest heavily on Theorem 4 in Sun et al. (1999), who considered models with densities $f_i(y_i \mid \eta_i)$ for the observations y_i given a predictor η_i and predictors $\boldsymbol{\eta}= (\eta_1, ..., \eta_i, ..., \eta_n)$ given by

$$\boldsymbol{\eta} = \boldsymbol{V\gamma} + \boldsymbol{Z_1\beta_1} + \boldsymbol{\varepsilon},$$

with partially improper prior for $\boldsymbol{\beta_1}$, and individual–specific random effects $\boldsymbol{\varepsilon}= (\varepsilon_1, ..., \varepsilon_i, ..., \varepsilon_n)' \sim N(\mathbf{0}, \tau_0^2\boldsymbol{I})$. We extend their theorem in two directions: First, we allow for several random effects with different degree and type of smoothness priors, and, second, we do not necessarily assume that individual–specific random effects ε_i are included in the predictor.

We consider models with predictor

$$\boldsymbol{\eta} = \boldsymbol{V\gamma} + \boldsymbol{Z_1\beta_1} + ... + \boldsymbol{Z_m\beta_m} + \boldsymbol{Z_0\beta_0}, \tag{10}$$

where $\boldsymbol{\gamma}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m$ have priors as in (8) and (5). The term $\boldsymbol{Z}_0\boldsymbol{\beta}_0$ represents a random effect with a $n \times d_0$ design matrix $\boldsymbol{Z}_0$, with $\text{rank}(\boldsymbol{Z}_0) = d_0 = \dim(\boldsymbol{\beta}_0)$, and a (possibly partially improper) prior

$$p(\boldsymbol{\beta}_0) \propto \tau_0^{-r_0} \exp\left(-\frac{1}{2\tau_0^2}\boldsymbol{\beta}_0'\boldsymbol{K}_0\boldsymbol{\beta}_0\right), \tag{11}$$

with $r_0 = \text{rank}(\boldsymbol{K}_0)$, such that

$$d_0 \geq d_j, \quad r_0 \geq r_j, \quad j = 1, ..., m.$$

Setting $\boldsymbol{Z}_0 = \boldsymbol{I}$, $\boldsymbol{\beta}_0 = \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \tau_0^2\boldsymbol{I})$, the predictor (10) also covers the case of individual–specific random effects $\boldsymbol{Z}_0\boldsymbol{\beta}_0 = \boldsymbol{\varepsilon}$. In geoadditive models $\boldsymbol{Z}_0\boldsymbol{\beta}_0$ will usually represent a spatial effect with a MRF or kriging prior, or an unstructured spatial effect.

Lemma A2

Consider a generalized linear mixed model with observation densities $f_i(y_i \mid \eta_i)$, predictor (10), and priors (8), (5), (11). Suppose that (after a reordering of observations)

$$(*) \qquad \int f_i(y_i \mid \eta_i) d\eta_i < \infty$$

holds for observations $i = 1, ..., n^*$, and

$$(**) \qquad f_i(y_i \mid \eta_i) \leq M, \quad i = n^*+1, ..., n$$

holds for the remaining observations.

Denote the corresponding submatrices of $\boldsymbol{V}$, $\boldsymbol{Z}$ and $\boldsymbol{Z}_0$ by $\boldsymbol{V}^*$, $\boldsymbol{Z}^* = (\boldsymbol{Z}_1^*, ..., \boldsymbol{Z}_m^*)$, $\boldsymbol{Z}_0^*$, and assume:

(iii) $\operatorname{rank}(\boldsymbol{Z}_0^*) = d_0$,

the rank conditions (i) in Lemma A1 hold for $\boldsymbol{V}^*$, $\boldsymbol{Z}^*$,

condition (ii) in Lemma A1 holds with r_0 replacing n and SSE replaced by SSE^*.

Then the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \tau_0^2, ..., \tau_m^2 \mid \boldsymbol{y})$ is proper.

The following corollary is easier to check.

Corollary A2

Assume that conditions $(*)$, $(**)$ and the rank conditions for $\boldsymbol{V}^*$, $\boldsymbol{Z}^*$, $\boldsymbol{Z_0}^*$ in Lemma A2 hold, and that

$$r_0 - p - (d - r) > 0$$

with $d = d_0 + ... + d_m$, $r = r_0 + ... + r_m$, and

$$a_j > 0, \quad b_j > 0, \quad j = 0, ..., m$$

hold for the inverse Gamma priors (11).

Then the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \tau_0^2, ..., \tau_m^2 \mid \boldsymbol{y})$ is proper.

Proofs: We consider first the simpler case of individual–specific random effects $\boldsymbol{\beta_0} \equiv \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \tau_0^2 \boldsymbol{I})$. Using the one–to–one relation $\boldsymbol{\eta} = \boldsymbol{V\gamma} + \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}$ between $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$, we consider propriety of $p(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tau_0^2, \boldsymbol{\tau}^2 \mid \boldsymbol{y})$ instead of $p(\boldsymbol{\varepsilon}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tau_0^2, \boldsymbol{\tau}^2 \mid \boldsymbol{y})$. Proceeding as in Sun et al. (1998), one starts from

$$p(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tau_0^2, \boldsymbol{\tau}^2 \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) p(\boldsymbol{\beta}) p(\tau_0^2) p(\boldsymbol{\tau}^2).$$

Using $(**)$ and integrating out $\boldsymbol{\eta}^{**} = (\eta_{n^*+1}, ..., \eta_n)$, one arrives at

$$\begin{aligned} p(\boldsymbol{\eta}^*, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tau_0^2, \boldsymbol{\tau}^2 \mid \boldsymbol{y}) &\propto \prod_{i=1}^{n^*} f_i(y_i \mid \eta_i) \{ p(\boldsymbol{\eta}^* \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) p(\boldsymbol{\beta}) p(\tau_0^2) p(\boldsymbol{\tau}^2) \} \\ &\propto \prod_{i=1}^{n^*} f_i(y_i \mid \eta_i) \{ p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \tau_0^2, \boldsymbol{\tau}^2 \mid \boldsymbol{\eta}^*) \}. \end{aligned}$$

Applying Lemma A1 (or Corollary A1) to

$$\boldsymbol{\eta}^* = \boldsymbol{V}^* \boldsymbol{\gamma} + \boldsymbol{Z}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}^*, \quad \boldsymbol{\varepsilon}^* \sim N(\boldsymbol{0}, \tau_0^2 \boldsymbol{I}),$$

gives

$$p(\boldsymbol{\eta}^* \mid \boldsymbol{y}) \propto \prod_{i=1}^{n^*} f_i(y_i \mid \eta_i),$$

and propriety follows from $(*)$.

For the general case $\boldsymbol{\eta} = \boldsymbol{V\gamma} + \boldsymbol{Z\beta} + \boldsymbol{Z}_0 \boldsymbol{\beta}_0$, with prior (11) for $\boldsymbol{\beta}_0$, we first decompose $\boldsymbol{\beta}_0$ into a $(d_0 - r_0)$–dimensional subvector $\boldsymbol{\beta}_0^{fl}$ with flat prior $p(\boldsymbol{\beta}_0^{fl}) \equiv 1$ and a r_0–dimensional subvector $\boldsymbol{\beta}_0^{pr}$ with a proper prior $\boldsymbol{\beta}_0^{pr} \sim N(\boldsymbol{0}, \tau_0^2 \boldsymbol{I})$:

$$\boldsymbol{\beta}_0 = \boldsymbol{Z}_0^{fl} \boldsymbol{\beta}_0^{fl} + \boldsymbol{Z}_0^{pr} \boldsymbol{\beta}_0^{pr},$$

where the $d_0 \times (d_0 - r_0)$ matrix $\boldsymbol{Z}_0^{fl}$ contains a basis of the nullspace of $\boldsymbol{K}_0$. The matrix $\boldsymbol{Z}_0^{fl}$ is the identity vector $\mathbb{1}$ for P–splines with first–order random walk prior, Markov–random fields and 2d–P–splines with MRF prior for the coefficients. For P–splines with second–order random walk prior it is a two column matrix whose first column is the identity vector and the second column is composed of the (equidistant) knots of the spline.

The $d_0 \times r_0$ matrix $\boldsymbol{Z}_0^{pr}$ is given by

$$\boldsymbol{Z}_0^{pr} = \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1},$$

where $\boldsymbol{L} = \boldsymbol{S}'\boldsymbol{\Lambda}^{1/2}$ is obtained from the spectral decomposition $\boldsymbol{K}_0 = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}'$ of $\boldsymbol{K}_0$. It follows that

$$\boldsymbol{\beta}_0^{pr} \sim N(\boldsymbol{0}, \tau_0^2 \boldsymbol{I}).$$

Defining $\tilde{\boldsymbol{V}} = (\boldsymbol{V}, \boldsymbol{Z}_0\boldsymbol{Z}_0^{fl})$, $\tilde{\gamma}' = (\boldsymbol{\gamma}, \boldsymbol{\beta}_0^{fl})'$, $\tilde{\boldsymbol{Z}}_0 = \boldsymbol{Z}_0\boldsymbol{Z}_0^{pr}$, we can rewrite the predictor as

$$\boldsymbol{\eta} = \tilde{\boldsymbol{V}}\tilde{\gamma} + \boldsymbol{Z}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}_0\boldsymbol{\beta}_0^{pr}.$$

For identifiability reasons, the columns of $\boldsymbol{Z}_0\boldsymbol{Z}_0^{fl}$ are not contained in the $(d_0 - r_0)$ column space of $\boldsymbol{V}$, so that $\text{rank}(\tilde{\boldsymbol{V}}) = p + (d_0 - r_0)$. Defining $\boldsymbol{\varepsilon}_0 = \tilde{\boldsymbol{Z}}_0\boldsymbol{\beta}_0^{pr}$, we have an additive mixed model

$$\boldsymbol{\eta} = \tilde{\boldsymbol{V}}\tilde{\gamma} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0 \tag{12}$$

for the predictor $\boldsymbol{\eta}$, with singular covariance matrix $\text{cov}(\boldsymbol{\varepsilon}_0) = \tilde{\boldsymbol{Z}}_0\tilde{\boldsymbol{Z}}_0'\tau_0^2$ of the 'error term' $\boldsymbol{\varepsilon}_0$. Let

$$\tilde{\boldsymbol{S}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{S}}' = \tilde{\boldsymbol{Z}}_0\tilde{\boldsymbol{Z}}_0'$$

be the spectral decomposition of $\tilde{\boldsymbol{Z}}_0\tilde{\boldsymbol{Z}}_0'$, with $\tilde{\boldsymbol{\Lambda}} = \text{diag}(\lambda_1, ..., \lambda_{r_0})$ containing the r_0 positive eigenvalues, and set

$$\boldsymbol{T} = \tilde{\boldsymbol{\Lambda}}^{-1/2}\tilde{\boldsymbol{S}}'.$$

Multiplying equation (12) by $\boldsymbol{T}$, we obtain the reduced model

$$\tilde{\boldsymbol{\eta}} = \boldsymbol{T}\tilde{\boldsymbol{V}}\tilde{\gamma} + \boldsymbol{T}\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \tau_0^2\boldsymbol{I}),$$

where $\tilde{\boldsymbol{\eta}} = \boldsymbol{T}\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon} = \boldsymbol{T}\boldsymbol{\varepsilon}_0$ have dimension r_0.

Altogether, we obtain a (linear) one–to–one transformation between $\boldsymbol{\beta}_0^{pr}$ and $\tilde{\boldsymbol{\eta}}$, and proving propriety of $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\tau}^2, \tau_0^2 \mid \boldsymbol{y})$ is equivalent to proving propriety of $p(\tilde{\boldsymbol{\eta}}, \tilde{\gamma}, \boldsymbol{\beta}, \boldsymbol{\tau}^2, \tau_0^2 \mid \boldsymbol{y})$.

Thus, we can repeat the arguments of the first part of the proof, replacing $\boldsymbol{\eta}$ by $\tilde{\boldsymbol{\eta}}$, $\boldsymbol{V}$ by $\boldsymbol{T}\tilde{\boldsymbol{V}}$, $\boldsymbol{Z}$ by $\boldsymbol{T}\boldsymbol{Z}$, and n by r_0.

From Magnus, Neudecker (1991, p. 273) it follows that

$$\text{rank}(\boldsymbol{T}\tilde{\boldsymbol{V}}) = \text{rank}(\tilde{\boldsymbol{V}}) = p + d_0 - r_0,$$

$$\text{rank}(\boldsymbol{T}\boldsymbol{Z}_j) = \text{rank}(\boldsymbol{Z}_j) = r_j.$$

Applying now Lemma A1 (or Corollary A1) to the model for $\tilde{\boldsymbol{\eta}}$, we obtain Lemma A2 and Corollary A2.

Proofs of Theorem 1 and Corollary 1: We first show that the conditions $(*)$, $(**)$ of Lemma A2 are fulfilled for right–censored survival data (t_i =min$(T_i, U_i), \delta_i)$, $\quad i = 1, ..., n$. The density of observation i is given by

$$f_i(t_i \mid \eta_i(t_i)) = \lambda_i(t_i)^{\delta_i} S_i(t_i),$$

where

$$\lambda_i(t_i) = \exp(\eta_i(t_i)), \quad S_i(t_i) = \exp\left(-\int_0^{t_i} \lambda_i(s)ds\right)$$

For censored observations ($\delta_i = 0$), we have $f_i(t_i \mid \eta_i(t_i)) = S_i(t_i) \leq 1$, so that condition $(**)$ of Lemma A2 holds.

For uncensored observations ($\delta_i = 1$)

$$f_i(t_i \mid \eta_i(t_i)) = \lambda_i(t_i)S_i(t_i).$$

Setting $\eta_i := \eta_i(t_i), \quad \lambda_i := \lambda_i(t_i)$, we obtain

$$\int_0^\infty f_i(t_i \mid \eta_i)d\eta_i = \int_0^\infty \lambda_i S_i(t_i)\lambda_i^{-1}d\lambda_i = \int_0^\infty S_i(t_i)d\lambda_i,$$

so that assumption $(*)$ is equivalent to

$$\int_0^\infty S_i(t_i)d\lambda_i < \infty. \tag{13}$$

We factorize the multiplicative hazard rate $\lambda_i(t)$ into

$$\lambda_i(t) = c_i l_i(t),$$

where $c_i > 0$ is the time–constant part. Then

$$\int_0^\infty S_i(t_i)d\lambda_i = \int_0^\infty \exp\left\{-c_i \int_0^{t_i} l_i(s)ds\right\} d\lambda_i.$$

Consider first the case where $\eta_i(t)$ is piecewise constant (on the intervals I_k, $k = 1, 2, ...$ defined by the knots of B–splines of degree 0). Then

$$\lambda_i(t) = c_i\lambda_{ik} \text{ for } t \in I_k,\ k = 1, 2, ...$$

For $t_i \in I_k$, say, we have $\lambda_i = \lambda_i(t_i) = c_i\lambda_{ik}$, and

$$\begin{aligned} \int_0^\infty S_i(t_i)d\lambda_i \quad &\propto \quad \int_0^\infty \exp\left(-\left(c_i \sum_{j=1}^{k-1} \Delta_j \lambda_{ij}\right) - c_i \int_{\xi_{k-1}}^{t_i} \lambda_{ik} d\lambda_{ik}\right) d\lambda_{ik} \\ &\propto \quad C_i \int_0^\infty \exp(-c_i(t_i - \xi_{k-1})\lambda_{ik})d\lambda_{ik} < \infty, \end{aligned}$$

for $t_i - \xi_{k-1} > 0$, which is valid a.s. for continuous T_i.

Consider now the case, where the time–varying part of $\eta_i(t)$ is defined by B–splines of higher degree. Let

$$\lambda_{ik} = \min_{t \in I_k} l_i(t) > 0, \quad k = 1, 2, ...$$

be the minimum of the time–varying part of $\lambda_i(t)$ on I_k.

Then

$$\int_0^\infty \exp\left\{-c_i \int_0^{t_i} l_i(s)ds\right\} d\lambda_i \leq C_i \int_0^\infty \exp\left\{-c_i \int_{\xi_{k-1}}^{t_i} \lambda_{ik} d\lambda_{ik}\right\} d\lambda_{ik}$$
$$= C_i \int_0^\infty \exp(-c_i(t_i - \xi_{k-1})\lambda_{ik}) d\lambda_{ik} < \infty,$$

so that assumption (13) is fulfilled. Corollary 1 immediately follows from Corollary 2.

Remark: We have tacitly made the assumption that $\lambda_i(t) > 0$ for any choice of covariates and parameters. This is valid because of our parametrization

$$\lambda_i(t) = \exp(\eta_i(t)).$$

References

Gelman, A. (2004), "Prior distributions for variance parameters in hierarchical models," provided by *Economics Working Paper Archive at WUSTL* in its series *Econometrics* with number 0404001.

Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2005), "Geoadditive Survival Models," *Revised for JASA*.

Magnus, J.R. and Neudecker, H. (1991), *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Statistics.

Speckman, P.L. and Sun, D. (2003), "Fully Bayesian spline smoothing and intrinsic autoregressive priors," *Biometrika*, 90, 2, 289–302.

Sun, D., Tsutakawa, R.K. and Speckman, P.L. (1999), "Posterior distribution of hierarchical models using CAR(1) distributions," *Biometrika*, 86, 341–350.