Fahrmeir, Kneib:

Propriety of Posteriors in Structured Additive Regression Models: Theory and Empirical Evidence

Sonderforschungsbereich 386, Paper 510 (2006)

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

Projektpartner

Propriety of Posteriors in Structured Additive Regression Models: Theory and Empirical Evidence

Ludwig Fahrmeir

Department of Statistics
Ludwig-Maximilians-University Munich
ludwig.fahrmeir@stat.uni-muenchen.de

Thomas Kneib

Department of Statistics
Ludwig-Maximilians-University Munich
thomas.kneib@stat.uni-muenchen.de

Abstract

Structured additive regression comprises many semiparametric regression models such as generalized additive (mixed) models, geoadditive models, and hazard regression models within a unified framework. In a Bayesian formulation, nonparametric functions, spatial effects and further model components are specified in terms of multivariate Gaussian priors for high-dimensional vectors of regression coefficients. For several model terms, such as penalised splines or Markov random fields, these Gaussian prior distributions involve rank-deficient precision matrices, yielding partially improper priors. Moreover, hyperpriors for the variances (corresponding to inverse smoothing parameters) may also be specified as improper, e.g. corresponding to Jeffery's prior or a flat prior for the standard deviation. Hence, propriety of the joint posterior is a crucial issue for full Bayesian inference in particular if based on Markov chain Monte Carlo simulations. We establish theoretical results providing sufficient (and sometimes necessary) conditions for propriety and provide empirical evidence through several accompanying simulation studies.

Key words: Bayesian semiparametric regression, Markov random fields, MCMC, penalised splines, propriety of posteriors

1 Introduction

Bayesian structured additive regression (STAR) has been proposed in Fahrmeir, Kneib & Lang (2004) as a comprehensive class of semiparametric regression models with continuous or discrete responses and different types of covariates and corresponding effects. Popular subclasses are generalized additive models, additive mixed models, and geoadditive models that consist of nonparametric effects of continuous covariates, spatial effects and cluster-specific random effects in different combinations. STAR models allow to combine these different model classes and a number of extensions in a unifying framework that also facilitates development of generally applicable inferential schemes. The same model class can be extended to the analysis of continuous survival times in structured hazard regression models (Hennerfeind, Brezger & Fahrmeir 2006).

A Bayesian formulation of STAR models involves specification of high-dimensional Gaussian smoothing priors for nonparametric functions, spatial effects and further model components. Typically, nonparametric functions are specified through Bayesian penalised splines (P-splines) with partially improper random walk priors for the B-spline coefficients. Priors for spatial effects can be formulated as stationary Gaussian random fields or Gaussian Markov random fields. While the former lead to proper Gaussian smoothing priors, the latter are again partially improper. In addition, priors for the variances of the smoothness priors (corresponding to inverse smoothing parameters) are frequently assumed to follow weakly informative inverse gamma distributions or limiting cases corresponding to flat, improper priors for variances or standard deviations. Full Bayesian inference, described in Fahrmeir et al. (2004) and Brezger & Lang (2006) for exponential family models and Hennerfeind et al. (2006) for hazard regression models, is based on Markov chain Monte Carlo (MCMC) simulations building upon sequential sampling from full conditional distributions. Since these full conditionals may be proper distributions even in the case of a non-existing, improper joint posterior, the crucial question is: Is the resulting joint posterior prior despite the (partially) improper formulation of some of the priors?

In this article, we present theorems guaranteeing propriety under certain assumptions related mainly to the hyperparameters of the inverse gamma priors of the variances and the rank deficiency of the precision matrices of the Gaussian smoothness priors. In addition,

we investigate performance of the MCMC algorithms in interesting limiting cases where, from a theoretical perspective, the joint posterior is still proper but close to an improper posterior. Furthermore, we provide some evidence that MCMC works well in some situations not covered by the (sufficient but not necessary) assumptions for propriety in the theorems.

Propriety of posteriors when priors are partially improper has been considered in various statistical models in the literature. Our theoretical results are mainly based on and extend important research by Sun, Tsutakawa & He (2001) and Speckman & Sun (2003) on propriety of posteriors in mixed models. However, the assumption of proper Gaussian smoothing priors in the former articles prevents direct application to STAR models. Sun & Speckman (2006) present results on propriety in Gaussian additive models build upon smoothing splines with partially improper priors but their results rely on properties specific to smoothing splines which are not applicable in the more general setting of STAR models.

To make results for usual mixed models applicable to STAR models, we make use of the mixed model representation of STAR models, which has been introduced in Fahrmeir et al. (2004) as a computational tool for empirical Bayes inference. The mixed model representation allows to rewrite STAR models as variance components mixed models with proper Gaussian priors. This allows to extend results presented in Sun et al. (2001) and Speckman & Sun (2003) to (the reparameterised) Gaussian STAR models or exponential family models with individual-specific random effects. Since such individual-specific effects can not be included in any exponential family regression model (e.g. binary models), we will introduce a further reparameterisation step that allows to overcome the necessity of individual-specific effects. We will also discuss how conditions formulated at the different stages of the reparameterised model relate to the original STAR model formulation. In a further step, we extend own work on propriety of Bayesian geoadditive survival models presented in Hennerfeind et al. (2006). Therefore we will again make use of the mixed model formulation introduced in Kneib & Fahrmeir (2007) for hazard regression models.

The paper proceeds as follows: Section 2 reviews basic STAR methodology and establishes the mixed model representation. Section 3 describes propriety in Gaussian STAR models and provides foundations for the more general model classes discussed in Section 4 for responses from exponential families and hazard regression models. The accompanying

simulation studies have been carried out with BayesX (Brezger, Kneib & Lang 2005), a software package that provides implementations of the discussed STAR models. BayesX is freely available from `http://www.stat.uni-muenchen.de/~bayesx`.

2 Structured Additive Regression

2.1 Exponential Family Models

2.1.1 Observation Model

Generalized linear models relate the expectation of response variables from exponential families to a linear predictor $\eta_i = u_i'\gamma$ formed by covariates u_i and regression coefficients γ via $E(y_i|u_i) = h(\eta_i)$, where h is a suitable known response function. To account for non-linear effects of continuous covariates, spatial correlations, unobserved heterogeneity or further non-standard covariate effects, several extensions of the basic linear model have been considered in the literature. A fairly general geoadditive mixed model is given by the predictor

$$\eta_i = u_i'\gamma + f_1(x_{i1}) + \ldots + f_k(x_{ik}) + f_{geo}(s_i) + b_{g_i}, \tag{1}$$

where $f_1(x_1), \ldots, f_k(x_k)$ are smooth functions of continuous covariates, $f_{geo}(s)$ is a spatial function defined upon either spatial coordinates $s = (s_x, s_y)$ or a discrete spatial lattice index $s \in \{s_1, \ldots, s_S\}$, and b_g is a cluster-specific random effect with grouping structure represented by the factor variable $g \in \{1, \ldots, G\}$. Geoadditive mixed models are a special case of a larger class of regression models called structured additive regression (STAR, Fahrmeir et al. (2004)) that attempts to combine different types of non-standard covariate effects in a unified framework. In addition to the model terms in Equation (1), STAR models may comprise random slopes $u_j b_{jg}$, interaction surfaces $f_{j,k}(x_j, x_k)$, and varying coefficient terms $u_j f(x_k)$ with continuous effect modifier x_k or $u_j f_{j,geo}(s)$ with spatial effect modifier s (see Fahrmeir et al. (2004) for a detailed description). In generic notation and after appropriate reindexing, a general STAR model can be described by the predictor

$$\eta_i = u_i'\gamma + f_1(v_{i1}) + \ldots + f_r(v_{ir}), \tag{2}$$

where a function $f_j(v_j)$ represents any of the effects discussed before and v_j is a generic covariate, which may be continuous, bivariate, or a spatial or grouping indicator depending

on the corresponding effect. Note that in general model (2) is not identifiable if no additional assumptions are made about the levels of some of the functions f_j. While no restrictions have to be imposed on varying coefficient terms, the remaining effects are usually assumed to be appropriately centered. In addition, an intercept term is included in the parametric part $u'\gamma$ to account for the overall level of the predictor. We will come back to the identifiability problem in the next sections where we discuss STAR models and prior assumptions in more detail.

A special case of STAR models are models with individual-specific random effects $b_{g_i} = b_i$, i.e. $\{1, \ldots, G\} = \{1, \ldots, n\}$. In this case, conditions for the propriety of posteriors can be formulated based on work of Sun et al. (2001). Note that Gaussian models are also included in this framework if the error terms ε_i are identified with individual-specific random effects, although the error variables are of course not parameters of interest. However, some models such as the binary logit model do not allow for the inclusion of individual-specific effects, since these are not identifiable from the data. Moreover, the general inclusion of individual-specific effects even in models where they are formally identifiable is usually not justified and such effects should only be included when they are required from a statistical modelling perspective. Therefore it is important to generalize results for models with subject-specific effects to reduced models without such effects. We will further pursue this issue in Section 4.

2.1.2 Model Components and Priors

All types of effects considered in STAR models can be expressed as the product of a suitably chosen design matrix V_j and and a (possibly large) vector of regression coefficients ξ_j. Accordingly, predictor (2) can be represented in matrix notation as

$$\eta = U\gamma + V_1\xi_1 + \ldots + V_r\xi_r, \tag{3}$$

where U is the usual design matrix of fixed effects. In a Bayesian framework, model formulation is completed by assigning appropriate priors to the function f_j or, in the predictor (3), the corresponding regression coefficients. In STAR models, these priors can be expressed in the generic form of a multivariate Gaussian distribution, i.e.

$$p(\xi_j) \propto \frac{1}{(\tau_j)^{\mathrm{rk}(K_j)/2}} \exp\left(-\frac{1}{2\tau_j}\xi_j' K_j \xi_j\right). \tag{4}$$

The precision matrix K_j plays the role of a penalty matrix and, depending on the model term at hand, penalizes large differences between adjacent parameters or large deviations from a global mean. In general, the precision matrix does not have full rank, i.e. $k_j = \text{rk}(K_j) \leq \dim(\xi_j) = d_j$. The rank deficiency represents the fact, that for most effects a specific part of f_j remains unpenalized. The amount of smoothness caused by the penalty is controlled by the variance parameter τ_j which can be interpreted analogously to the smoothing parameter in a frequentist setting of nonparametric regression. Large values of τ_j allow for a strong variation in the regression coefficients ξ_j corresponding to wiggly function estimates, while a reverse implication holds for small variances.

In order to obtain identifiable STAR models, the prior distributions of some of the effects have to augmented by appropriate centering restrictions. This can be achieved by putting certain linear restrictions on the coefficients ξ_j which effectively reduces the dimension of ξ_j and the rank-deficiency of K_j by one for the corresponding effects (compare Rue & Held (2005) for a detailed discussion on priors of the form (4) subject to linear restrictions). While the specific form of appropriate restrictions is difficult to specify in the original model formulation, the mixed model representation of STAR models discussed in Section 2.3 leads to easy and interpretable indentifiability restrictions.

In a full Bayesian approach, the variance parameters τ_j are considered as hyperparameters which have to be estimated jointly with the remaining effects. The conjugate hyperprior to the multivariate Gaussian prior (4) is of the inverse Gamma type $\tau_j \sim IG(a_j, c_j)$ with density

$$p(\tau_j) \propto \frac{1}{(\tau_j)^{a_j+1}} \exp\left(-\frac{c_j}{\tau_j}\right). \tag{5}$$

For positive values $a_j > 0$ and $c_j > 0$ the prior is proper, while improper priors result for either $a_j \leq 0$ or $c_j \leq 0$. By allowing for improper priors, Equation (5) contains several special cases of particular interest:

- Setting $a_j = -1$ and $c_j = 0$ corresponds to a flat prior for the variance τ_j, i.e. $p(\tau_j) \propto \text{const}$.
- Setting $a_j = -0.5$ and $c_j = 0$ corresponds to a flat prior for the standard deviation $\sqrt{\tau_j}$, i.e. $p(\sqrt{\tau_j}) \propto \text{const}$.
- Setting $a_j = c_j = 0$ results in Jeffrey‘s prior, i.e. $p(\tau_j) \propto 1/\tau_j$.

To make the generic representation of STAR models more intuitive, we will now discuss some special cases in more detail. For smooth effects of continuous covariates and as a building block of varying coefficient terms, penalized splines have proven to be a valuable tool (see Eilers & Marx (1996) for a frequentist and Brezger & Lang (2006) for a Bayesian description of penalised splines). The basic principle is to approximate a function $f_j(x_j)$ by a linear combination of d_j basis functions, i.e.

$$f_j(x_j) = \sum_{m=1}^{d_j} \xi_{jm} B_m(x_j). \tag{6}$$

The design matrix V_j then consists of the basis functions evaluated at the observed covariate values (i.e. $V_j[i,m] = B_m(x_{ij})$) while the amplitudes ξ_{jm} are collected in the coefficient vector ξ_j. For varying coefficient terms, each row of the design matrix has to be multiplied by the value of the interaction variable in addition. When B-spline basis functions are employed in Equation (6), the prior for ξ_j is usually constructed based on random walks of order q_j, e.g.

$$\xi_{jm} = \xi_{j,m-1} + u_{jm} \qquad \text{or} \qquad \xi_{jm} = 2\xi_{j,m-1} - \xi_{j,m-2} + u_{jm}$$

in case of first and second order random walks with Gaussian error terms $u_{jm} \sim N(0, \tau_j)$. This leads to a penalty matrix $K_j = D_j' D_j$ formed by the crossproduct of a q_j-th order difference matrix D_j. Correspondingly, a $(q_j - 1)$-th order polynomial remains unpenalized by the precision matrix and prior (4) is partially improper with $\text{rk}(K_j) = d_j - q_j$. An alternative representation of P-splines is given by a truncated power series basis expansion with ridge penalty on the coefficients of the truncated basis functions (Ruppert, Wand & Carroll 2003). Again a polynomial (represented by the untruncated polynomials in the basis) remains unpenalized and the precision matrix for the full coefficient vector is rank deficient.

Similar ideas can be employed for modelling interaction surfaces $f_{j,k}(x_j, x_k)$ by defining bivariate basis functions, e.g. based on Tensor products of the univariate bases in x_j and x_k direction. Correspondingly, the penalty concept has to be adapted and a bivariate random walk may be considered. As for univariate splines the design matrix is constructed from evaluations of the basis functions and the penalty matrix is defined via Kronecker products of difference matrices for the univariate bases. Thus, the resulting penalty matrix for the bivariate effect is also rank-deficient, possibly with a higher dimensional null space resulting from interactions of the null spaces in x_j and x_k direction.

For geographical effects $f_{geo}(s)$ with a spatial lattice index $s \in \{s_1, \ldots, s_S\}$ Markov random field (MRF) priors are a suitable choice. MRFs extend commonly known temporal random walk priors to the spatial case of two-dimensional irregular lattices leading (in the simplest case) to the following prior for $\xi_s = f_{geo}(s)$:

$$\xi_s | \xi_{s'}, s' \neq s, \tau_{geo} \sim N\left(\frac{1}{N_s}\sum_{s' \in \partial_s} \xi_{s'}, \frac{\tau_{geo}}{N_s}\right), \tag{7}$$

where δ_s consists of the neighbors of index s and $N_s = |\delta_s|$ denotes the number of such neighbors. Computing the joint distribution of the vector $\xi_{geo} = (\xi_{s_1}, \ldots, \xi_{s_S})'$ yields a distribution of the form (4), where the precision matrix is given by an adjacency matrix, compare Rue & Held (2005, Ch. 3) for details. Since rows and columns in the adjacency matrix sum to zero, the precision matrix has a rank-deficiency of one and prior (7) is therefore also called an intrinsic MRF. For spatial coordinates $s = (s_x, s_y)$, spatial effects can be included as in traditional geostatistical models by assuming a zero mean Gaussian process with variance τ for $\{\xi_s, s \in \mathbb{R}^2\}$. In case of a finite set of coordinates, the joint distribution of all ξ_s is again multivariate Gaussian with the inverse correlation matrix as precision matrix K_j Obviously, the precision matrix is of full rank in this case. A compact description of the correlation structure is achieved by assuming a parametric correlation function for the Gaussian process, e.g. a member of the Matérn family. In both approaches to spatial modelling the design matrix is simply a 0/1 incidence matrix linking observations with the corresponding entries in the vector ξ_{geo}, i.e. $V_{geo}[i, s]$ equals one when observation i is located at site or coordinate s and zero otherwise.

As a last special case of (4) consider i.i.d. Gaussian random effects with respect to a grouping indicator $g \in \{1, \ldots, G\}$. In this case the joint distribution is a proper multivariate Gaussian distribution with precision (and correlation) matrix $K_j = I_G$. Similar as for the spatial effect, observations and random effects are linked by an incidence matrix as design matrix V_j.

2.1.3 Posterior and Sampling Scheme

The joint posterior of all effects in a STAR model is obtained using Bayes' Theorem as

$$p(\xi_1, \ldots, \xi_r, \tau_1, \ldots, \tau_r, \gamma | y) \propto L(y, \xi_1, \ldots, \xi_r, \gamma) \prod_{j=1}^{r} \left[p(\beta_j | \tau_j) p(\tau_j)\right],$$

where $L(.)$ denotes the likelihood derived from the exponential family assumption for the response. An efficient sampling scheme for STAR models can now be constructed based on Metropolis Hastings steps for the regression coefficients and Gibbs sampling steps for the variances. More precisely, we consider an iteratively weighted least squares (IWLS) proposal for ξ_j based on a Gaussian approximation to the full conditional with precision matrix and mean

$$P_j = V_j'WV_j + \frac{1}{\tau_j}K_j \qquad \text{and} \qquad m_j = P_j^{-1}V_j'W(\tilde{y} - \eta_{-j}),$$

where the diagonal matrix W and the vector of working observations $\tilde{y}$ are constructed in complete analogy to the usual GLM case (compare Fahrmeir & Tutz (2001)) and $\eta_{-j} = \eta - V_j\xi_j$ denotes the predictor without the j-th effect. Similar expressions are obtained for the vector of fixed effects, compare Brezger & Lang (2006) for details. The full conditional of τ_j is inverse Gamma with updated parameters

$$a_j' = a_j + \frac{1}{2}\,\mathrm{rk}(K_j) \qquad \text{and} \qquad c_j' = c_j + \frac{1}{2}\xi_j'K_j\xi_j.$$

Note that in general the full conditionals of the parameter blocks are all proper distributions, although the joint posterior may be improper. In particular, it is often not possible to determine impropriety of the posterior from the output of the MCMC simulation.

2.2 Structured Hazard Regression

Similar extensions as considered in Section 2.1 for exponential family regression can also be defined for hazard regression models when analysing survival data (t_i, δ_i), where t_i is an observed duration time and δ_i is the usual censoring indicator for right censored durations (compare Hennerfeind et al. (2006)). A geoadditive model comparable to (1) is given by a hazard rate $\lambda_i(t) = \exp(\eta_i(t))$ with

$$\eta_i(t) = h_0(t) + z_{i1}h_1(t) + \ldots + z_{il}h_l(t) + f_1(x_{i1}) + \ldots + f_k(x_{ik}) + b_{g_i} + u_i'\gamma. \qquad (8)$$

In addition to the geoadditive effects already discussed in the previous section, the predictor (8) contains an expression for the log-baseline hazard $h_0(t) = \log(\lambda_0(t))$ and several time-varying effects $h_l(t)$ of covariates z_l. Due to the inclusion of time-varying effects, structured hazard regression models are not restricted to the assumption of proportional hazards. Of course, similar extensions of the geoadditive model as mentioned in Section 2.1.1 can be considered in the survival case.

In generic notation, the predictor of a structured hazard regression model can also be expressed in the form (2), where the generic functions $f_j(v_j)$ may now also be time-dependent when representing a time-varying effect. Both the log-baseline hazard and time-varying effects can be modelled using penalised splines for a representation of $h_j(t)$, $j = 0, \ldots, l$. In particular, time-varying effects can be subsumed in the varying coefficient framework, if the survival time is considered the effect modifier. The posterior of structured hazard regression models and an MCMC sampling scheme can be derived in a similar form as in Section 2.1.3, compare Hennerfeind et al. (2006) for details.

2.3 Mixed Model Representation

In the following we will introduce a general mixed model representation of both structured additive models within the exponential family framework and structured hazard regression for continuous survival times. The fact that many penalisation approaches are equivalent to specific mixed models has received considerable attention throughout recent years and has been used to estimate semiparametric regression models in a variety of settings (compare Ruppert et al. (2003) for an overview, Fahrmeir et al. (2004) for results on exponential family STAR models, and Kneib & Fahrmeir (2007) for mixed model based hazard regression). In addition, the mixed model representation allows to adapt conditions for proper posteriors in mixed models to the more general case of STAR models. Sun & Speckman (2006) employed the mixed model representation of smoothing splines to derive conditions for purely additive models consisting of several smooth effects. However, the conditions presented in Sun & Speckman (2006) do only apply to models with Gaussian responses and are furthermore restricted to purely additive smoothing spline models. Semiparametric models which are usually required in most applications are not supported since the propriety conditions rely heavily on properties of smoothing splines. In Sections 3 and 4 we will therefore extend the more general conditions presented in Sun et al. (2001) to STAR models.

To rewrite STAR models as mixed models, consider a model term $V_j\xi_j$ with $\text{rk}(K_j) = k_j < d_j = \dim(\xi_j)$. For model terms with proper priors no reparametrisation is needed since in this case ξ_j can be directly interpreted as a (generally correlated) random effect. Applying a general result for partially improper Gaussian distributions (see Rue & Held (2005), p. 91), allows to partition ξ_j into a $(d_j - k_j)$-dimensional vector of fixed effects β_j

with improper prior and a k_j-dimensional vector of random effects b_j with proper prior. More specifically, ξ_j is decomposed into two parts as

$$\xi_j = \tilde{X}_j \beta_j + \tilde{Z}_j b_j, \tag{9}$$

where $\tilde{X}_j \beta_j$ captures the part of ξ_j which is unpenalized by K_j and $\tilde{Z}_j b_j$ captures the orthogonal deviation from the unpenalized part. Correspondingly, the design matrices $\tilde{X}_j$ and $\tilde{Z}_j$ can be constructed from the eigen decomposition of the penalty matrix K_j. The design matrix of β_j consists of the eigen vectors corresponding to the zero eigen values thereby representing a basis of the null space of K_j. The design matrix of b_j can then be constructed from the remaining eigen vectors corresponding to positive eigen values, compare Fahrmeir et al. (2004) for details. Choosing appropriate design matrices in (9) leads to the interpretation of β_j as a vector of fixed effects with noninformative prior $p(\beta_j) \propto$ const and b_j as a vector of i.i.d. random effects with Gaussian prior $b_j|\tau_j \sim N(0, \tau_j)$. The advantage of partition (9) is the explicit differentiation between an improper and a proper part which are mixed in a complex manner in the original prior (4).

Inserting the partition of ξ_j into the representation of a vector of function evaluations yields

$$V_j \xi_j = V_j(\tilde{X}_j \beta_j + \tilde{Z}_j b_j) = X_j \beta_j + Z_j b_j.$$

Collecting the indices of all model terms with partially improper priors in the set $\mathcal{J} \subset \{1, \ldots, r\}$ and the indices of model terms with proper priors in $\bar{\mathcal{J}} = \{1, \ldots, r\} \setminus \mathcal{J}$ finally allows to rewrite any structured additive predictor as

$$\begin{aligned} \eta &= U\gamma + V_1\xi_1 + \ldots + V_r\xi_r \\ &= U\gamma + \textstyle\sum_{j \in \mathcal{J}} X_j \beta_j + \sum_{j \in \bar{\mathcal{J}}} V_j \xi_j + \sum_{j \in \mathcal{J}} Z_j b_j \\ &= X\beta + Zb, \end{aligned} \tag{10}$$

where $X = (U, X_j, j \in \mathcal{J})$, $\beta = (\gamma', \beta_j', j \in \mathcal{J})'$, $Z = (V_j, j \in \bar{\mathcal{J}}, Z_j, j \in \mathcal{J})$ and $b = (\xi_j, j \in \bar{\mathcal{J}}, b_j', j \in \mathcal{J})'$.

Remark 1. In order to obtain a full rank design matrix of fixed effects X, superfluous columns constructed in the reparametrisation have to be deleted from X. These superfluous columns arise from the non-identifiability of the level for some of the functions f_j as discussed in Section 2.1.1. For these functions the design matrix X_j contains a column of ones modelling the overall level. Deleting this column is an easy and interpretable way to

include the centering restriction and is equivalent to the assumption that the corresponding regression coefficient in the vector β_j is set to zero. Using the one-to-one relationship in (9), we can also deduce the corresponding linear restriction on the original coefficient vector ξ_j. In the following sections we will always assume that the design matrix X has full rank, i.e. appropriate centering restrictions have been imposed on the regression coefficients ξ_j and no overparameterised models are considered.

3 Propriety in Gaussian STAR Models

The basic idea to obtain conditions for the propriety of the posterior distribution in Gaussian STAR models

$$y = U\gamma + V_1\xi_1 + \ldots + V_r\xi_r + \varepsilon, \qquad \varepsilon \sim N(0, \tau_0 I) \tag{11}$$

is to rewrite the original model in mixed model representation (10) with proper priors for the random effects b. For the variance parameter τ_0 we assume an additional inverse Gamma-type prior with hyperparameters a_0 and c_0. As discussed in the previous section, the original matrix of fixed effects U (with full rank p) is augmented to the matrix X in the mixed model representation. Let q denote the number of linear independent columns augmented to U after deletion of superfluous columns. Then the resulting matrix X has full rank $p + q$.

Define $V = (V_1, \ldots, V_r)$ and $\xi = (\xi_1', \ldots, \xi_r')'$. Then model (11) can be written as

$$y = U\gamma + V\xi + \varepsilon = W\theta + \varepsilon$$

where $W = (U, V)$ and $\theta = (\gamma', \xi')$. Let t be such that

$$\mathrm{rk}(U, V) = \mathrm{rk}(W) = p + t$$

and

$$SSE = y'(I - W(W'W)^{-}W')y$$

the usual sum of squares for the linear model $y = W\theta + \varepsilon$, where $(W'W)^{-}$ is a generalized inverse of $W'W$.

The following conditions together with Theorem 1 extend Theorem 2 of Sun et al. (2001) to Gaussian STAR models with partially improper priors for random effects:

(a) For $j = 1, \ldots, r$, either

(a1) $a < c_j = 0$, or

(a2) $c_j > 0$ hold.

(b1) $k_j + 2a_j > 0, \quad j = 1, \ldots, r.$

(b2) $k_j + 2a_j > \sum_{j=1}^{r} k_j - t + q, \quad j = 1, \ldots, r.$

(c1) $n - p - q + 2a_0 + 2(a_1 + \ldots a_r) > 0.$

(c2) $n - p - q + 2a_0^- + 2(a_1^- + \ldots + a_n^-) > 0$, where $a_j^- = \min(0, a_j)$.

Note that (c1) and (c2) are identical if $a_j < 0$ for $j = 0, \ldots, r$.

Theorem 1. *Consider the Gaussian STAR model (11) with* $\mathrm{rk}(U) = p$ *and assume that* $SSE + 2c_0 > 0$.

1. *If* $t = k + q$ *or if* $r = 1$ *the conditions (a), (b2), and (c1) are necessary, and conditions (a), (b2), and (c2) are sufficient for the propriety of the joint posterior.*

2. *If* $t < k + q$ *and* $r > 1$, *conditions (a), (b1), and (c1) are necessary, and conditions (a), (b2) and (c2) are sufficient for the propriety of the joint posterior.*

Remark 2. The assumption $SSE + 2c_0 > 0$ is obviously always fulfilled for $c_0 > 0$. In the case of an improper inverse Gamma-type prior with $a_0 < 0$, $c_0 = 0$, we have to assure that $SSE > 0$. If the number $\dim(\theta) = \dim(\gamma) + \dim(\xi)$ of parameters is equal or larger than n, the data $y = (y_1, \ldots, y_n)'$ can be interpolated by the predictor, so that $SSE = 0$. For $n > \dim(\theta)$, which will hold in many applications, we have $SSE > 0$ (almost surely) and we can choose $c_0 = 0$.

Theorem 1 covers some special choices for the Gamma-type priors which are of particular interest in practical work.

(i) (ϵ, ϵ)-priors:

Setting $a_j = c_j = \varepsilon$ for some small $\varepsilon > 0$ leads to the so-called proper inverse Gamma $(\varepsilon, \varepsilon)$-priors. These priors have been quite popular in applied work because they seem to circumvent well known problems with the limiting case $\varepsilon = 0$, of Jeffrey's prior. Note that for this limiting case the necessary condition (a) is not

fulfilled. Although these $(\varepsilon, \varepsilon)$-priors lead to proper posteriors under the simplified conditions of Corollary 1, there is some debate in the literature about sensitivity of posteriors and numerical stability in practical work, see for example Lambert et al. (2005) and Gelman (2006). We provide some empirical evidence for STAR models through a simulation study.

(ii) Flat priors for standard deviations $\sqrt{\tau_j}$:

The prior $p(\tau_j) \propto \tau_j^{-1/2}$ is equivalent to the choice $p(\sqrt{\tau_j}) \propto const$ for standard deviations and corresponds to an improper Gamma-type prior with hyperparameters $a_j = -1/2$, $c_j = 0$. This prior corresponds to a special case of (aa), and is recommended as a standard choice in practical work by Gelman (2006).

(iii) Flat priors for variances τ_j:

A flat prior $p(\tau_j) \propto const$ is an improper inverse Gamma-type prior with $a_j = -1$, $c_j = 0$. This prior corresponds to REML estimation of variance parameters τ_j in an empirical Bayes approach to STAR models, see Fahrmeir et al. (2004). More exactly, the REML estimate can be interpreted as a posterior mode estimate for τ_j, while full Bayesian inference via MCMC provides the posterior mean estimate. Corollary 1 summarizes propriety conditions for these special cases.

Corollary 1. Consider a Gaussian STAR model (11) with $\mathrm{rk}(U) = p$. The following conditions are (jointly) sufficient (and partly necessary) for propriety of posteriors in the cases (i), (ii), and (iii):

(i) (ϵ, ϵ) - priors:

$$n > p + q.$$

(ii) Flat priors for standard deviations:

$$SSE > 0,$$
$$k_i > k + q + 1 - t, \quad i = 1, \ldots, r, \quad \text{where } k = k_1 + \ldots + k_r, \text{ and}$$
$$n > p + q + r + 1.$$

(iii) Flat priors for variances:

$$SSE > 0,$$

$$k_i > k + q + 2 - t, \quad i = 1, \ldots, r, \text{ and}$$

$$n > p + q + r + 2.$$

Proof of Theorem 1. The proof is basically an application of Theorem 2 in Sun et al. (2001) to the reparametrized model (10), where X is the design matrix for unpenalized effects β after deleting superfluous columns constructed in the reparametrization. Then X has full rank $\tilde{p} = p + q$. It is easy to see that

$$\text{rk}(U, V) = \text{rk}(X, Z),$$

implying $\text{rk}(X, Z) = p + t$. Because X has full rank $\tilde{p} = p + q$, it can be shown that

$$\text{rk}(X, Z) = p + t \Leftrightarrow \text{rk}(RZ) = \tilde{t} := t - q$$

where $R = I - X(X'X)^{-1}X'$.

Now Theorem 2 of Sun et al. can be applied to the mixed model (10), replacing p and t in their conditions through $\tilde{p} = p + q$ and $\tilde{t} = t - q$, respectively. Note also that $SSE > 0$ in the original STAR model implies $SSE_{re} > 0$ in the reparametrized model.

Corollary 1 follows immediately from Theorem 1 as a special case.

∇

Remark 3. Propriety in terms of the original (γ, ξ)-parameterisation follows from the one-to-one relationship (9), including the centering restrictions as explained in Remark 1.

To provide some empirical evidence on the theoretical results derived in Theorem 1, we conducted some simulation studies investigating different aspects of Theorem 1. First of all, we focused on the (ϵ, ϵ)-type priors. Therefore, we simulated 100 replications of the Gaussian nonparametric model $y_i = \sin(x_i) + \varepsilon_i$, $i = 1, \ldots, 50$ with ϵ_i i.i.d. $N(0, 0.4^2)$ and estimated the model with a cubic P-spline with second order random walk penalty and 20 inner knots in combination with various specifications for the hyperparameters of the variances. To be more specific, we considered $a_j = c_j = 0.001$, $a_j = c_j = 0.0001$, $a_j = c_j = 0.00001$, $a_j = c_j = 0.000001$, and $a_j = c_j = 0$. For comparison, we also included the flat priors for variances $(a_j = -1, c_j = 0)$ as well as the standard deviations $(a_j = -0.5, c_j = 0)$. Figure 1 displays the corresponding empirical log-MSEs

$$\log\left(\frac{1}{50}\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))\right).$$

Obviously, the choice of hyperparameters has hardly any influence on the results, in particular when both hyperparameters are equal and small. To gain more insight into the impact of hyperparameter settings, Figure 2 shows sampling paths for some of the hyperparameters combinations and one particular simulation run. To obtain comparable results, the simulations have been started with the same seed. The sampling paths mostly confirm results obtained from the consideration of log-MSEs. The differences between different (ϵ, ϵ)-priors are almost invisible. Both types of flat priors yield a somewhat increased variability in the sampling paths which is more expressed in case of a flat prior for the variance.

As a key conclusion, it seems to make hardly a difference whatever value is specified for ε. Even in the case $\varepsilon = 0$, where the posterior is improper according to Theorem 1, results still coincide with those obtained with, say, $\varepsilon = 0.001$.

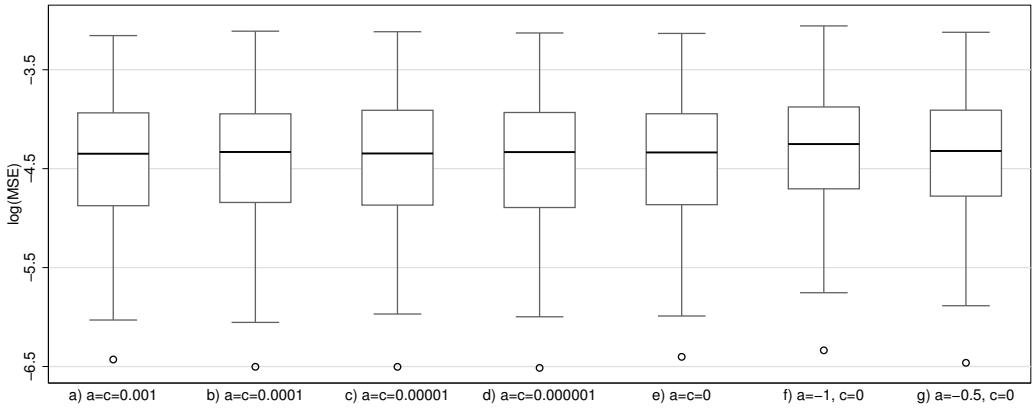

Figure 1: Gaussian nonparametric model: Boxplots of log-MSEs for $\hat{f}(x)$ for various specifications of hyperparameters a and b.

In a second simulation we aimed at investigating the restriction $SSE + 2c_0 > 0$. Therefore, we set up the geoadditive model $y_i = \sin(x_i) + f_{spat}(s_i) + \varepsilon_i$, $i = 1, \ldots, 124$ with ε_i i.i.d. $N(0, 0.4^2)$ and a spatial function $f_{spat}(s)$ defined upon the 124 districts of the southern part of Germany (Bavaria and Baden-Württemberg). The nonparametric function is again modelled by cubic P-Spline with 20 inner knots and second order random walk. The spatial effect is assigned a Markov random field prior. In total, the model contains more parameters than observations and, as a consequence, it is possible that $SSE = 0$ due to an interpolating fit. Note however, that the quantity SSE considered in Theorem 1 does not account for the effective dimension reduction introduced by the penalty terms.

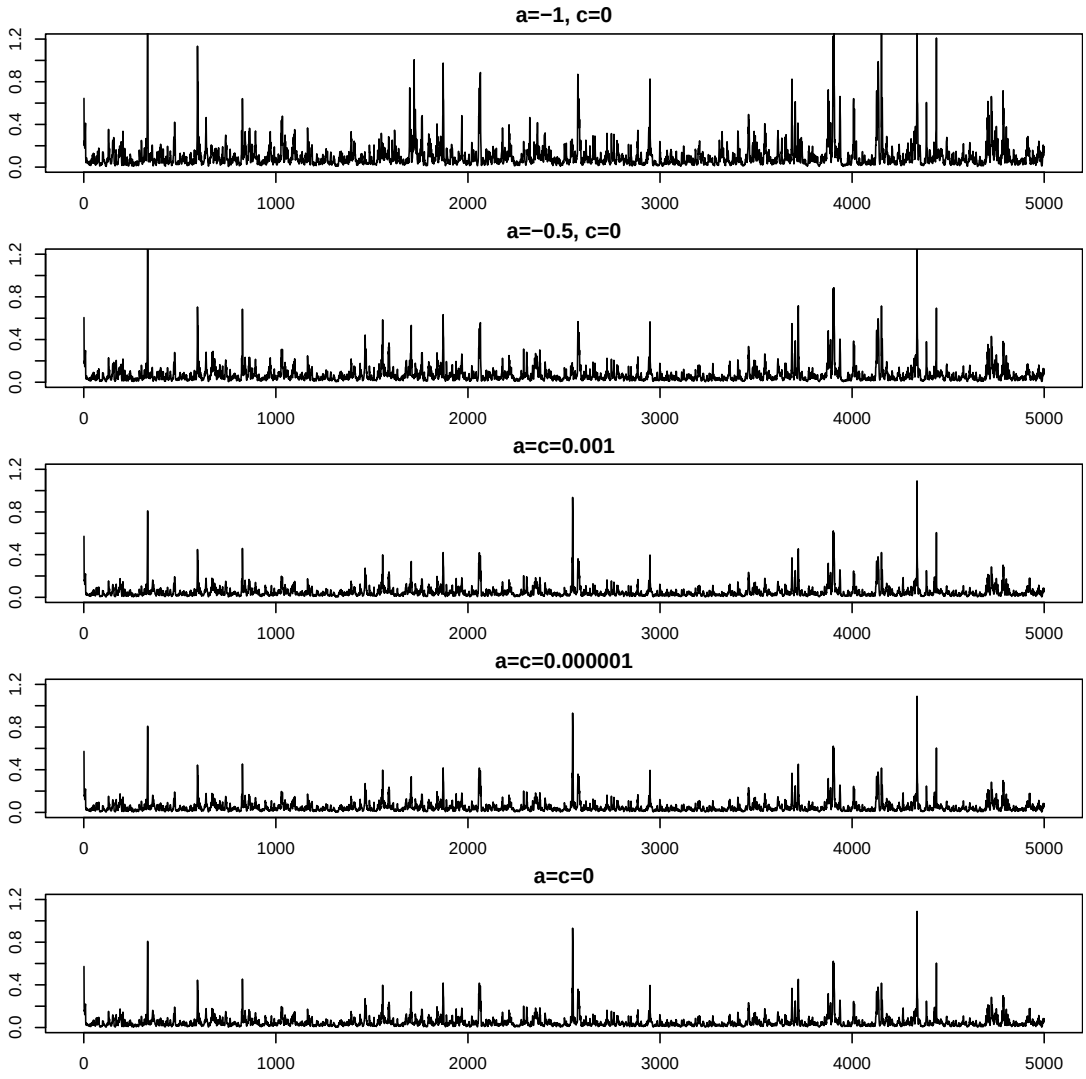

Figure 2: Gaussian nonparametric model: Sampling paths for the variance of the smooth function $f(x)$ for various specifications of hyperparameters a and b.

Hence, the fitted model itself does not suffer from overfitting due to interpolation.

We simulated 100 replications of the model and estimated them with both types of lat priors ($a_j = -1, c_j = 0$ and $a_j = -0.5, c_j = 0$) as well as with versions with a slightly positive $c_j = 0.001$. While in the former case, the assumption $SSE + 2c_0 > 0$ may be violated, the latter ensures $SSE + 2c_0 > 0$. Figure 3 display sampling paths for the variance component of the spatial effect for the different hyperparameter specifications. Visually there is no difference between the cases with $c_j = 0$ and $c_j = 0.001$. Also, differences between flat priors for the variances and flat priors for the standard deviations are only moderate. This is also confirmed by the estimated spatial effects. Figure 4 displays average estimates obtained from the 100 simulation runs. All effects are very

close to each other as well as to the true underlying function. Similar results are obtained for the nonparametric effect (results not shown).

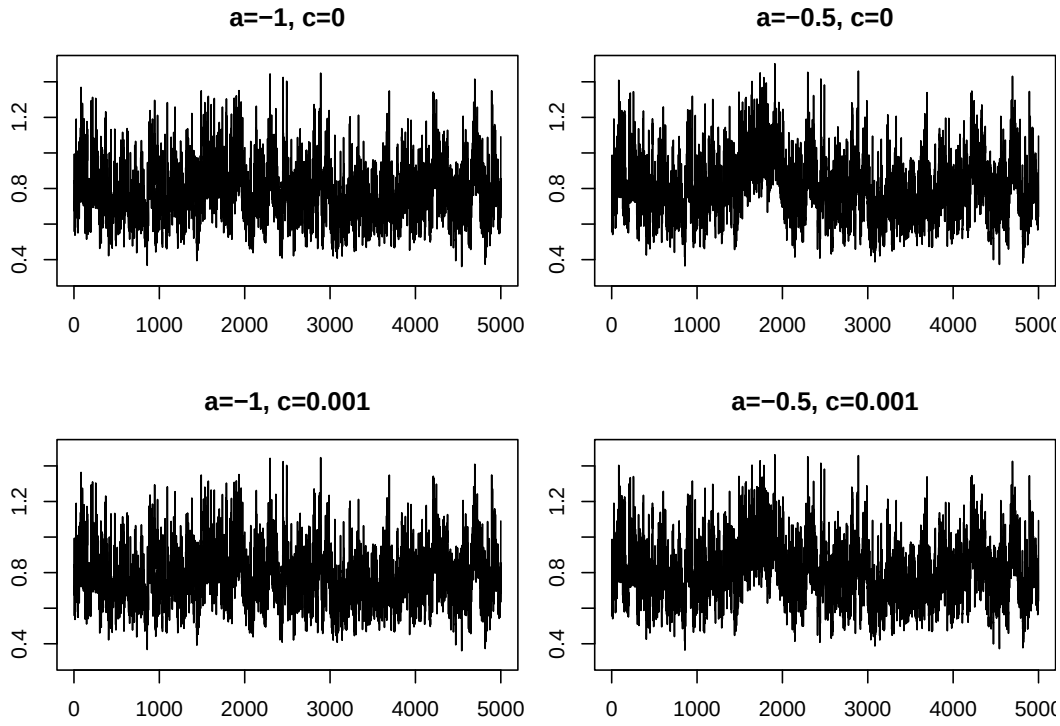

Figure 3: Gaussian geoadditive model: Sampling paths for the variance of the spatial effect $f_{spat}(s)$ for various specifications of hyperparameters a and b.

4 Propriety in Non-Gaussian STAR Models

4.1 Exponential Family Models

This section deals with STAR models where the (conditional) distribution of the response is a member of the univariate exponential family. We focus on models without an additional dispersion parameter, including binary, binomial and Poisson STAR models as the most important special cases. Extensions to models with additional dispersion parameter such as negative binomial or gamma models are briefly discussed at the end of the section.

We first consider one-parameter models with densities $f_i(y_i|\eta_i)$ for conditionally independent observations y_i given a predictor $\eta_i, i = 1, \ldots, n$, and predictors $\eta = (\eta_1, \ldots, \eta_n)$

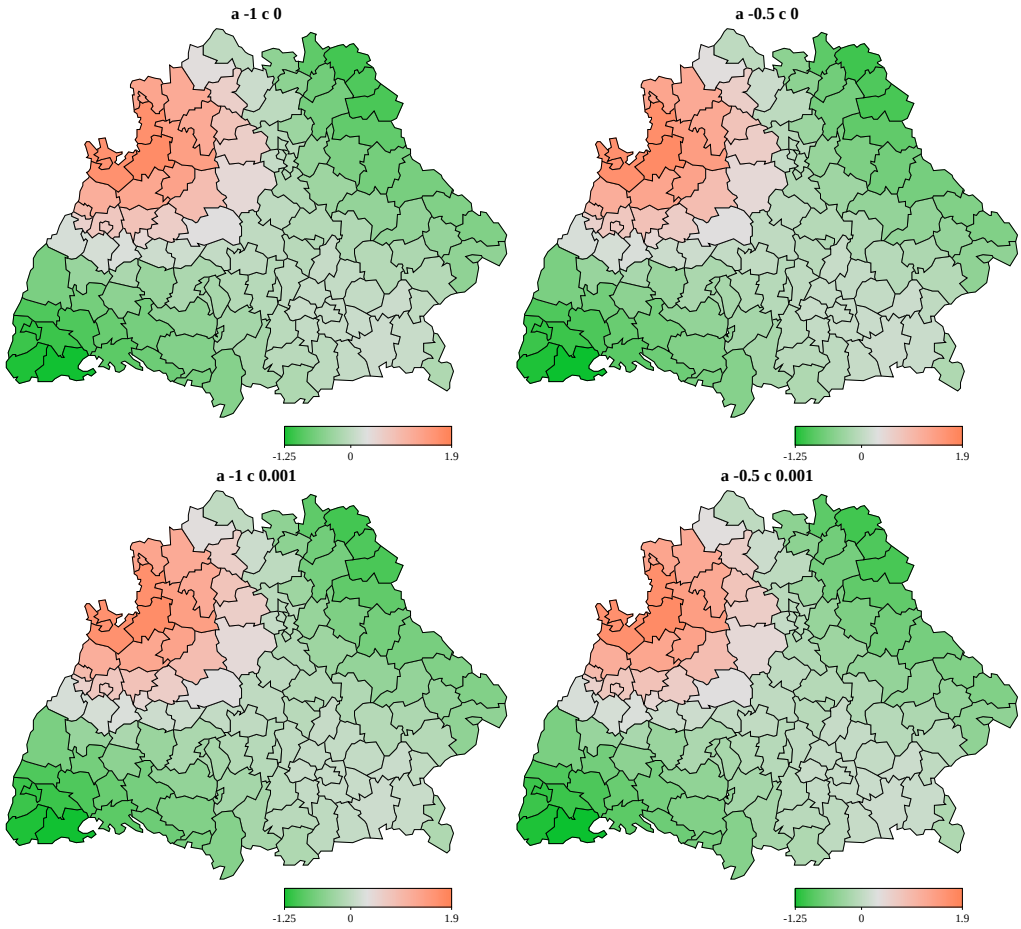

Figure 4: Gaussian geoadditive model: Average of the estimated spatial effects $\hat{f}_{spat}(s)$ for various specifications of hyperparameters a and b.

given by

$$\eta = U\gamma + V_1\xi_1 + \ldots + V_r\xi_r + V_0\xi_0, \tag{12}$$

where $\gamma, \xi_1, \ldots, \xi_r$ have the same priors $p(\gamma) \propto$ const and (4) as in the Gaussian case. The additional term $V_0\xi_0$ represents a random effect with full rank $n \times d_0$ design matrix V_0, i.e. $\text{rk}(V_0) = d_0 = \dim(\xi_0)$, and a (possibly partially improper) prior of the form (4) for ξ_0, i.e.

$$p(\xi_0) \propto \frac{1}{(\tau_0)^{k_0/2}} \exp(-\frac{1}{2\tau_0}\xi_0' K_0 \xi_0),$$

with $k_0 = \text{rk}(K_0)$, such that

$$d_0 \geq d_j, \qquad k_0 \geq k_j, \qquad j = 1, \ldots, r.$$

The variance parameter τ_0 is assumed to have a Gamma-type prior (5) with hyperparameters a_0, c_0.

Setting $V_0 = I, \xi_0 = \varepsilon \sim N(0, \tau_0 I)$, the predictor (12) also covers the case of individual-specific random effects $V_0\xi_0 = \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. In geoadditive models, $V_0\xi_0$ will usually represent a spatial effect with a MRF or kriging prior, or an unstructured spatial effect. In generalized additive models, $V_0\xi_0$ will represent the penalized spline with the largest number of basis functions or knots.

A mixed model representation as in Section 2.3, including the additional term $V_0\xi_0$, is

$$\eta = X\beta + Zb + Z_0 b_0, \qquad b_0 \sim N(0, \tau_0 I) \tag{13}$$

with $\dim(b_0) = k_0$. The augmented design matrix X may possibly contain additional columns constructed from the unpenalized part of ξ_0. Again q is the number of additional columns, augmenting U to X, such that X has full rank $p + q$.

The basic idea to obtain propriety results is to transform model (13) to a model $\tilde{\eta} = \tilde{X}\gamma + \tilde{Z}b + \varepsilon, \varepsilon \sim N(0, \tau_0 I)$ with reduced dimension $\dim(\tilde{\eta}) = \dim(\tilde{\varepsilon}) = k_0$. This model has random effects b with proper priors and "individual-specific" effects ε, so that ideas and results in Sun et al. can be applied.

We make the following Assumptions (i) - (iv): After a reordering of observations

(i) $\int f_i(y_i|\eta_i)d\eta_i < \infty$ holds for observations $i = 1, \ldots, n^*$, and

(ii) $f_i(y_i|\eta_i) \leq M$ holds for the remaining observations $i = n^* + 1, \ldots, n$.

Denoting the submatrices of $U, X, V = (V_1, \ldots, V_r)$ and V_0 corresponding to $i = 1, \ldots, n^*$ by U^*, X^*, V^* and V_0^*, we assume the rank conditions

(iii) $\mathrm{rk}(U) = \mathrm{rk}(U^*) = p$, $\mathrm{rk}(X) = \mathrm{rk}(X^*) = p + q$, $\mathrm{rk}(U, V) = \mathrm{rk}(U^*, V^*) = p + t$, and $\mathrm{rk}(V_0^*) \geq k_0$.

The rank condition for V_0^* allows to select k_0 linear independent rows from V_0^*, corresponding to a selected set $\{i_1, \ldots, i_{k_0}\} \subset \{1, \ldots, n^*\}$ of observations. We denote the corresponding submodel by

$$\eta_s = U_s\gamma + V_s\xi + V_{0s}\xi_0, \tag{14}$$

where U_s, X_s, U_s, X_s and X_{0s} denote corresponding submatrices. We further assume that

(iv) $\mathrm{rk}(X_s) = \mathrm{rk}(X^*) = p + q$ and $\mathrm{rk}(U_s, V_s) = p + \tilde{t}$, where $\tilde{t} \leq t$.

Remark 4.

1. The condition $\text{rk}(V_0^*) \geq k_0$ implies that (14) with predictor η_s is a submodel of the model defined by the observations $i = 1, \ldots, n^*$. This simplifies the proof of Theorem 2, using arguments from the proof of Theorem 3 in Sun et al. It can be omitted however, as we outline in a comment after Theorem 2.

2. Conditions (i) and (ii) correspond to conditions (B1), (B2) in Sun et al. (2001), who assume individual-specific random effects $V_0\xi_0 =: \varepsilon \sim N(0, \tau_0 I)$ in the predictor η. In this case, condition (iv) can be omitted.

Theorem 2. *Consider an exponential family STAR model with predictor (12). Assume that conditions (i) to (iv) and conditions (a), (b2) and (c2) in Theorem 1 hold with k_0 replacing n and $\tilde{t}$ replacing t. If $SSE_s + 2c_0 > 0$, where SSE_s is the usual residual sum of squares of the submodel (14), then the joint posterior is proper.*

Proof of Theorem 2. We rewrite the submodel (14) in mixed model representation as

$$\eta_s = X_s\beta + Z_s b + \varepsilon_s \tag{15}$$

where $\varepsilon_s = Z_{0s}b_0 \sim N(0, \tau_0 W)$, and $W = Z_{0s}Z_{0s}'$ has full rank k_0. Multiplication by $W^{-1/2}$ from the left leads to the normalized model

$$v = \tilde{X}\beta + \tilde{Z}b + \varepsilon, \qquad \varepsilon \sim N(0, \tau_0 I). \tag{16}$$

Because $W^{-1/2}$ is nonsingular, conditions (iv) for the submodel (14) and (15) also hold for the normalized model (16).

We show that the posterior $p(\beta, b, b_0, \tau, \tau_0|y)$ of the parameters in mixed model representation (16) is proper under the conditions of Theorem 2. Obviously, this is equivalent to prove that the posterior $p(\beta, b, v, \tau, \tau_0|y)$ is proper. The proof of the latter statement can be based on ideas and results in the proof for Theorem 3 of Sun et al. (2001). From Bayes' Theorem we have

$$\begin{aligned} p(\beta, b, v, \tau, \tau_0|y) &\propto \prod_{i=1}^{n} f(y_i|\eta_i) p(v|\beta, b, \tau_0) p(b|\tau) p(\tau) p(\tau_0) \\ &\propto \prod_{i=1}^{n} f(y_i|\eta_i) \frac{1}{\tau_0^{k_0/2}} \exp\left(-\frac{1}{2\tau_0}(v - \tilde{X}\beta - \tilde{Z}b)'(v - \tilde{X}\beta - \tilde{Z}b)\right) \\ &\quad \times \frac{1}{|Q|^{1/2}} \exp\left(-\frac{b'Q^{-1}b}{2}\right) p(\tau) p(\tau_0), \end{aligned}$$

where $Q = \text{cov}(b)$. Defining

$$G = \frac{1}{\tau_0^{k_0/2}|A|^{1/2}} \exp\left(-\frac{1}{2\tau_0}(v - \tilde{X}\beta - \tilde{Z}b)'(v - \tilde{X}\beta - \tilde{Z}b) - \frac{b'A^{-1}b}{2}\right)$$

we get

$$p(\beta, b, v, \tau_0|y) \propto \prod_{i=1}^{n} f(y_i|\eta_i)G.$$

Assumption (ii) implies

$$p(\beta, b, v, \tau, \tau_0|y) \leq M^* \prod_{i=1}^{n^*} f(y_i|\eta_i)G,$$

where $M^* = M^{n-n^*}$, and integrating out β, b and τ gives

$$p(v, \tau_0|y) \leq M^* \prod_{i=1}^{n^*} f(y_i|\eta_i) \int G dbd\beta d\tau.$$

The integral corresponds to expression G_3 in(A.17) of Sun et al. (2001) (omitting integration over the additional parameters $\varrho_1, \ldots, \varrho_r$), and it can be bounded from above as in their expressions (A.25) and (A.27). Using the assumptions (a), (b2) and (c2), replacing n by k_0 and t by $\tilde{t}$, we get the inequality

$$p(v, \tau_0|y) \leq \tilde{M} \prod_{i=1}^{n^*} f(y_i|\eta_i) g(\tau_0),$$

where $\tilde{M}$ is a generic constant and

$$g(\tau_0) = \tau_0^{-(k_0-p-q)/2-a_0-(a_1^-+\ldots+a_r^-)-1} \exp\left\{-\frac{SSE_s + 2c_0}{2\tau_0}\right\}.$$

Assumption (c2), with k_0 replacing n, and $SSE_s + c_0 > 0$ imply $\int g(\tau_0)d\tau_0 < \infty$ and therefore

$$p(v|y) \leq C \prod_{i=1}^{n^*} f(y_i|\eta_i),$$

where C is a generic constant. The final step is to show that $\int p(v|y)dv < \infty$.

Using the relation $v = W^{-1/2}\eta_s$ between v in the normalized model (16) and η_s in the unnormalized model (15) we have

$$\int p(v|y)dv = \int p(\eta_s|y)|W^{1/2}|d\eta_s \leq K \int \prod_{i=1}^{n^*} f(y_i|\eta_i)d\eta_s$$

for some constant K. Assumptions (i) and (ii) now imply $\int p(v|y)dv < \infty$ and thus proprietry of the posterior $p(\beta, b, b_0, \tau, \tau_0|y)$.

∇

Remarks 5.

1. In condition (iii) we have assumed that $\text{rk}(V_0^*) \geq k_0$, guaranteeing that k_0 linear independent rows of V_0^* can be selected, so that V_{0s} in (14) and Z_{0s} in (14) have full rank k_0. We consider now the more general case, where $0 < \text{rk}(V_0^*) = k_0^* < k_0$. Then we can still select k_0^* linear independent rows from V_0^* as well from Z_0^*, and $k_0 - k^*$ rows from the remaining rows of V_0 and Z_0 corresponding to observations $i = n^* + 1, \ldots, n$. Let SSE_s^* denote the k_0^*-dimensional submodel which replaces (14) in this more general situation. A slight extension of the proof of Theorem 2 shows that the posterior is still proper under the remaining conditions, if k_0 is replaced by k_0^* and SSE_s by SSE_s^*.

2. Theorem 2 can be extended to exponential family models with additional nuisance parameter in similar manner as Theorem 4 of Sun et al. We omit the details here.

Similarly as for Gaussian response models, we conducted a simulation study for Bernoulli, Binomial (with five replications) and Poisson distributed responses to investigate propriety of posteriors under (ϵ, ϵ)-priors. For all response distributions the sample size was given by $n = 50$ and the predictor is $\eta_i = \sin(x_i)$. We applied models with natural link functions, i.e. the logit link for Bernoulli and Binomial distributed responses and the log-link for the Poisson model. Each simulation consisted of 100 replications and the function was estimated with a cubic P-spline with 20 inner knots and second order random walk prior. For the hyperparameters we considered the same specifications as for Gaussian responses, i.e. $a_j = c_j = 0.0001$, $a_j = c_j = 0.00001$, $a_j = c_j = 0.000001$, $a_j = c_j = 0$, $a_j = -1, c_j = 0$ and $a_j = -0.5, c_j = 0$. In the case of exponential family regression, we experienced the (expected) difficulties for the limiting case $a_j = c_j = 0$. As an illustration, Figure 5 shows a selected sampling path for the variance of the nonparametric effect in case of a Binomial distributed response. While both specifications with $a_j < 0$ and $c_j = 0$ yield sampling paths with at least some variability, this variability strongly decreases when considering (ϵ, ϵ)-priors. In the limiting case with $\epsilon = 0$ the variance remains constantly equal to zero for a long time, resulting in a full conditional which is no longer of the inverse gamma type, since $c_j' = c_j + 1/2\xi_j K_j \xi_j = 0$ (at least numerically). The same problem occurs for the two remaining response distributions (results not shown).

Figure 6 displays log-MSEs for the Binomial model and the hyperparameter specifications that resulted in a proper posterior. The results do not vary that much, but there is a slight tendency for an increased MSE in case of decreasing hyperparameter values. Both types

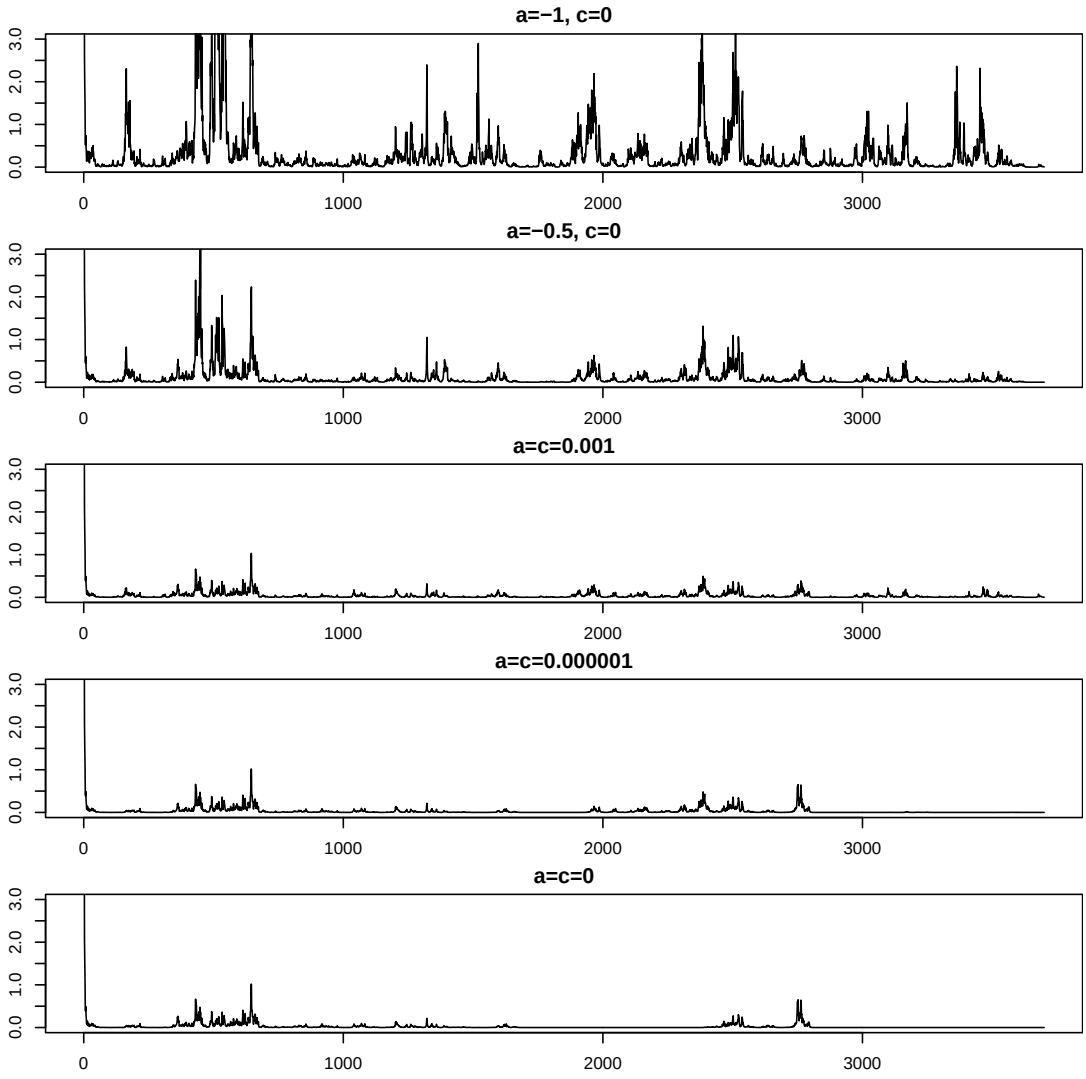

Figure 5: Binomial nonparametric model: Sampling paths for the variance of the smooth function $f(x)$ for various specifications of hyperparameters a and b.

of flat priors result in similar estimates, comparable to those from the $a_j = c_j = 0.001$ prior although the sampling paths in Figure 6 look quite different. Obviously, the function estimates do not respond very sensibly to the choice of hyperparameters.

In a second simulation, we again considered the question of whether $SSE_s + 2c_0$ is really required to obtain a proper posterior. In case of general exponential family regression this condition is even more restrictive than in the Gaussian case, since the sample size n is replaced by the sample size k_0 in the submodel. We considered additive models with $\eta_i = sin(x_{i1}) + x_{i2}^2$ where both effects are modelled as cubic P-splines with 20 inner knots and second order random walk prior. Hence, the sample size in the submodel equals the number of unknown parameters therefore allowing $SSE_s = 0$ due to interpolation.

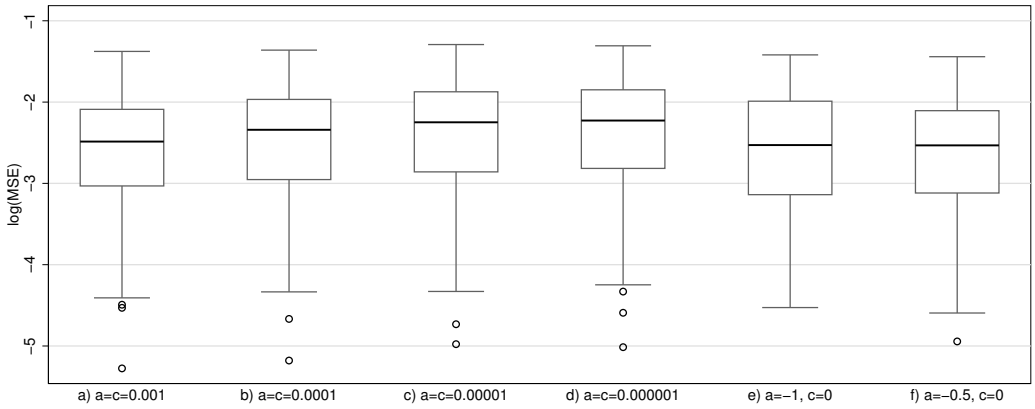

Figure 6: Binomial nonparametric model: Boxplots of log-MSEs for $\hat{f}(x)$ for various specifications of hyperparameters a and b.

Figure 7 shows sampling paths for the variance component of the sine curve when either $c_j = 0$ or $c_j = 0.001$ in case of Bernoulli distributed responses. For both types of flat priors, the sampling paths are visually indistinguishable with increased variability when $a_j = -1$. Figure 8 visualizes log-MSEs for the sine curve with the same sets of hyperparameters and an additional, intermediate version. Obviously, the results are again quite insensitive to the choice of hyperparameters for a fixed value of a_j. When comparing results with $a_j = -1$ and $a_j = -0.5$, there seems to be a slight improvement when considering the latter, i.e. a flat prior for the standard deviation. Finally, Figure 9 displays estimates for the quadratic effect averaged over the simulation runs as an exemplary result.

When considering Binomial and Poisson distributed results, the findings of the simulation study where qualitatively of the same type as for Bernoulli distributed response. We therefore decided not to present these results in detail.

4.2 Structured Hazard Regression

Propriety of posteriors in structured hazard regression models can be shown in a similar setup as for exponential family models but in this case the differentiation between observations in conditions (i) and (ii) in Section 4.1 is induced by the censoring of some of the observations. Let $\eta_i := \eta_i(t_i)$ denote the value of the predictor (8) at the observed lifetime t_i, $i = 1, \ldots, n$, and $\eta = (\eta_1, \ldots, \eta_n)'$ the predictor vector. Correspondingly, $g_0 = (g_0(t_1), \ldots, g_0(t_n))'$ is the vector of evaluations of the log-baseline hazard $g_0(t)$, and

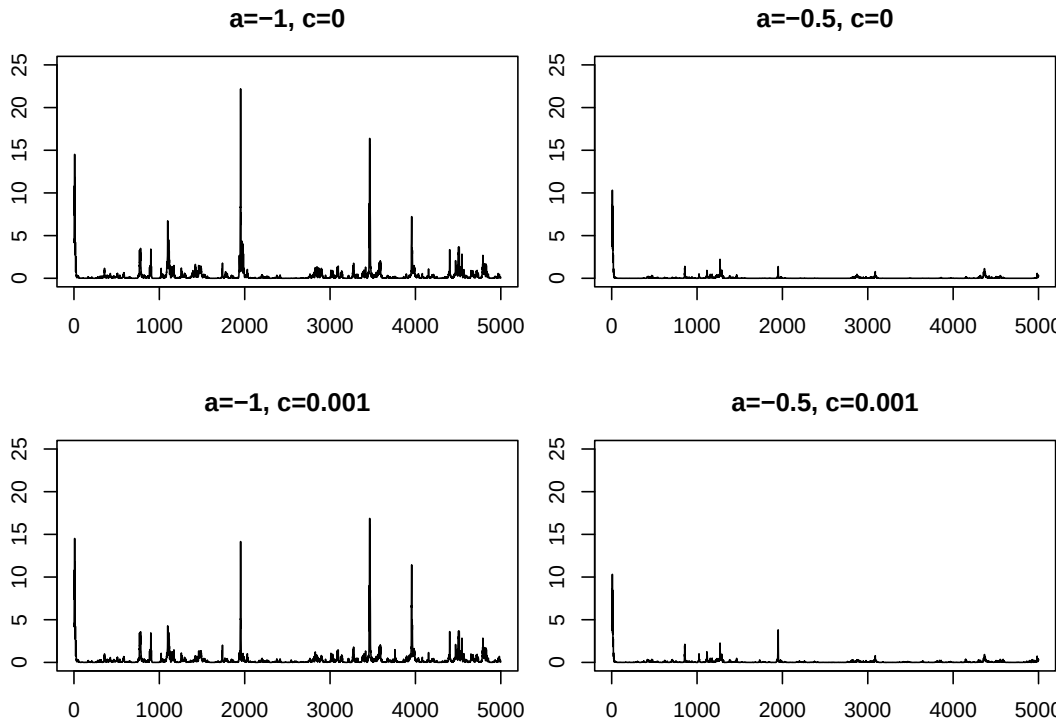

Figure 7: Binomial additive model: Sampling paths for the variance of the nonparametric effect $f_2(x_2)$ for various specifications of hyperparameters a and b.

$g_j = (g_j(t_1)z_{1j})', j = 1, \ldots, l$, the vectors of evaluations of the time-varying effect components in (8). Because the functions $g_j(t), j = 0, \ldots, l$, are modelled through Bayesian P-splines, the vectors g_j can be expressed in the generic form $g_j = V_j\xi_j$, with prior (4) for ξ_j, and the predictor η can be written in the form (12). The term $V_0\xi_0$ is defined as in Section 4.1, and will, for example, represent individual specific effects or a spatial effect.

Theorem 3. *Consider a hazard rate model with hazard rate $\lambda_i(t) = \exp(\eta_i(t))$ and structured additive predictor (8). Assume that, after reordering, observations $i = 1, \ldots, n^*$ are uncensored, and observations $i = n^*+1, \ldots, n$ are censored. If conditions (iii) and (iv) of Section 4.1 and the remaining conditions in Theorem 2 hold, then the joint posterior is proper.*

Proof of Theorem 3. We first note that the proof of Theorem 2 does not make use of the exponential family form of the densities $f(y_i|\eta_i)$ in conditions (i) and (ii). It still holds, if we replace $f(y_i|\eta_i)$ through the likelihood contribution of (right-censored) observed lifetimes $(t_i, \delta_i), i = 1, \ldots, n$, given by

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

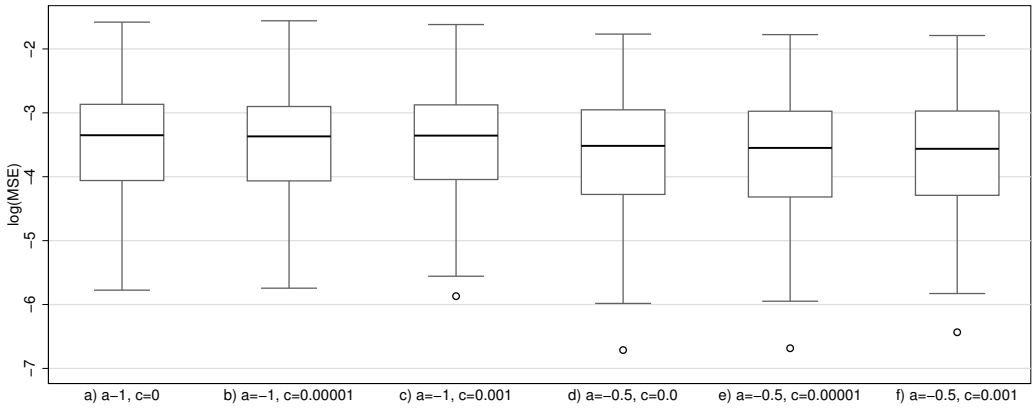

Figure 8: Binomial additive model: Boxplots of log-MSE for the nonparametric effect $f_2(x_2)$ and various specifications of hyperparameters a and b.

where

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

Therefore, we only have to show that conditions (i) and (ii) hold for uncensored and censored lifetimes, respectively. 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 (ii) holds.

For uncensored observations ($\delta_i = 1$) the likelihood contribution is given by

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

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

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

so that condition (i) is equivalent to

$$\int_0^\infty S_i(t_i)d\lambda_i < \infty. \tag{17}$$

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

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

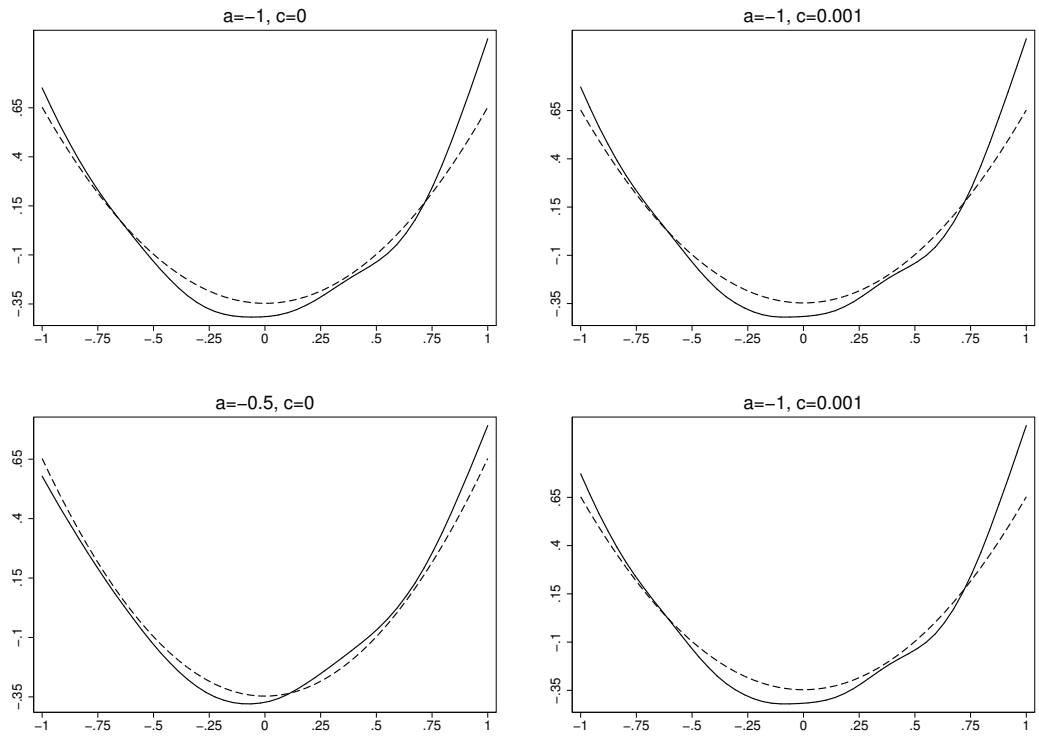

Figure 9: Binomial additive model: Average of the estimated nonparametric effects $\hat{f}_2(x_2)$ for various specifications of hyperparameters a and b. For comparison, the true effect is included as dashed line.

Consider first the case where all time-varying functions are represented by B-splines of degree 0, i.e. $\eta_i(t)$ is piecewise constant on the intervals I_k, $k = 1, 2, \ldots$ defined by the knots of the B-spline. 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

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

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

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

so that assumption (17) is fulfilled.

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

$\triangledown$

Remark 6. Similarly as for Gaussian and exponential family responses, we investigated the theoretical results in Theorem 3 through simulation studies. The results were qualitatively equivalent to those from Section 4.1.

5 Summary

In this paper, we developed necessary (and partly sufficient) theoretical conditions for propriety of posteriors in a large class of semiparametric regression models and supplemented these with results from several simulation studies. Based on a mixed model representation, results developed for mixed models could be applied to models with individual-specific random effects and Gaussian regression models. A further reparameterisation step allowed to formulate propriety conditions even in models without such individual-specific effects. We also made some attempts to trace back the porpriety conditions to the original formulation of STAR models to obtain a more intuitive interpretation. The performed simulation studies provided some empirical evidence that MCMC algorithms even work well in situations not covered by the (sufficient) conditions presented in the theorems, emphasizing the need for further research in the direction of sufficient *and* necessary conditions for propriety.

References

Brezger, A., Kneib, T. & Lang, S. (2005). BayesX: Analysing Bayesian structured additive regression models. *Journal of Statistical Software*, **14** (11).

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

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

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

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models*, Springer, New York.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, to appear.

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, to appear.

Kneib, T. & Fahrmeir, L. (2007) A Mixed Model Approach for Geoadditive Hazard Regression. *Scandinavian Journal of Statistics*, **??**, ??-??.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields. Theory and Applications.* CRC / Chapman & Hall, London.

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

Speckman, P. L. & Sun, D. (2003). Fully Bayesian Spline Smoothing and Intrinsic Autoregressive Priors. *Biometrika*, **90**, 289-302.

Sun, D. & Speckman, P. L. (2006). Bayesian Hierarchical Mixed Models for Additive Smoothing Splines. *Annals of Statistics*, to appear.

Sun, D., Tsutakawa, R. K. & He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica*, **11**, 77-95.