Final version published in
Computational Statistics and Data Analysis, 161: 107250.
doi.org/10.1016/j.csda.2021.107250

Fast Bayesian inference using Laplace approximations in nonparametric double additive location-scale models with right- and interval-censored data

Philippe Lambert

Institut de Recherche en Sciences Sociales (IRSS), Méthodes Quantitatives en Sciences Sociales, Université de Liège, Place des Orateurs 3, B-4000 Liège, Belgium

Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium.

Abstract

Penalized B-splines are commonly used in additive models to describe smooth changes in a response with quantitative covariates. This is usually done through the conditional mean in the exponential family using generalized additive models with an indirect impact on other conditional moments. Another common strategy is to focus on several low-order conditional moments, leaving the full conditional distribution unspecified. Alternatively, a multi-parameter distribution could be assumed for the response with several of its parameters jointly regressed on covariates using additive expressions. The latter proposal for a right- or interval-censored continuous response with a highly flexible and smooth nonparametric density is considered. The focus is on location-scale models with additive terms in the conditional mean and standard deviation. Starting from recent results in the Bayesian framework, a fast converging algorithm is proposed to select penalty parameters from their marginal posteriors. It is based on Laplace approximations of the conditional posterior of the spline parameters. Simulations suggest that the estimators obtained in this way have excellent frequentist properties and superior efficiencies compared to approaches with a working Gaussian assumption. The methodology is illustrated by the analysis of right- and interval-censored income data.

Keywords: Location-scale model ; Dispersion model; Imprecise data ; Interval-censoring ; P-splines ; Laplace approximation ; Constrained density estimation.

1. Introduction

Additive models are flexible alternatives to the classical linear regression model to describe in a flexible way the effect of quantitative covariates on various aspects of a response distribution. Early proposals focussed on the conditional mean with limited assumptions on the conditional distribution of the response (Breiman and Friedman, 1985). That idea was used to extend generalized linear models (GLM, Nelder and Wedderburn, 1972) and the analysis of nonnormal data (such as counts or proportions) in the framework of the exponential family of distributions: additive terms enter the GLM linear predictor (connecting covariates to a pre-specified

Email address: p.lambert@uliege.be (Philippe Lambert)
URL: www.statsoc.ulg.ac.be (Philippe Lambert)

function of the conditional mean) for a fixed value of the dispersion parameter, yielding generalized additive models (GAM) (Hastie and Tibshirani, 1986, 1990), see R-packages `mgcv` (Wood, 2017) and `blapsr` (Gressani and Lambert, 2021) for an implementation in a frequentist or fully Bayesian framework, respectively. Further extensions are possible by enabling covariates to also affect other aspects of the response distribution such as dispersion, skewness and kurtosis, see Lambert and Lindsey (1999) for early work on this with the four parameters of the stable distribution simultaneously modelled and Rigby and Stasinopoulos (2005) for an extension to a large choice of parametric distributions implemented in the R package `gamlss`. Lee et al. (2006, Chap. 11) and Gijbels and Prosdocimi (2012) considered joint additive models for location and dispersion within, respectively, the exponential and the double-exponential families of distributions, while Croux et al. (2012) relied on a (robustified) extended quasi-likelihood method.

The focus will be on double additive models for the conditional mean and standard deviation in location-scale models with a nonparametric error distribution. The response will be assumed continuous and possibly subject to right or interval-censoring. Nonparametric inference from censored data in location-scale models has been investigated by many authors, see e.g. Fan and Gijbels (1994) for early work using local polynomials and Heuchenne and Van Keilegom (2010) with the references therein for some more recent work. These methods typically focus on the estimation of the conditional location and can only handle the estimation of the smooth effects of a very limited number of covariates. Additive models based on P-splines (Eilers and Marx, 1996; Marx and Eilers, 1998; Lang and Brezger, 2004; Gressani and Lambert, 2021) are preferred here for their excellent properties (Eilers and Marx, 2010) and the possibility to handle a large number of additive terms. They are used to specify the joint effect of covariates on location and dispersion in the framework of the location-scale model, see Section 2. A nonparametric error distribution with an underlying smooth hazard function and fixed moments will be assumed for the standardized error term, see Section 2.5. In the absence of right-censoring, a location-scale model with a small number of additive terms and a quartile-constrained error density (instead of the hazard here) was considered in Lambert (2013) to analyse interval-censored data, with inference relying on a numerically demanding MCMC algorithm. It is shown how Laplace approximations to the conditional posterior of spline parameters can be combined to bring fast and reliable estimation of the linear and additive terms in the location and dispersion models, and provide a smooth estimate of the underlying error hazard function under moment constraints. These approximations are the cornerstones in the derivation of the marginal posteriors for the penalty parameters and smoothness selection, see Sections 2.4 and 2.5.5. The resulting estimation procedures are motivated using Bayesian arguments and shown to own excellent frequentist properties, see Section 3 and Appendix B. They are extremely fast and can handle a large number of additive terms within a few seconds even with pure R code. The methodology is illustrated in Section 4 with the analysis of right- and interval-censored income data in a survey. Section 5 closes the paper with a discussion and research perspectives.

2. Additive location-scale model

Consider a vector $(Y, \mathbf{z}, \mathbf{x})$ where Y is a univariate continuous response, $\mathbf{z}$ a $p-$vector of categorical covariates, and $\mathbf{x}$ a $J-$vector of quantitative covariates. The response could be subject to right-censoring, in which case one only observes (T, Δ), where $T = \min\{Y, C\}$, $\Delta = I(Y \leq C)$ and C denotes the right-censoring value that we shall assume independent of Y given the covariates. The response could also be interval-censored, meaning that it is only known to lie within an interval (Y^L, Y^U).

Such settings are common not only in survival analysis when studying the time elapsed between a clearly defined time origin and an event of interest, but also in surveys when the respondent reports a quantitative response by pointing an interval or a semi-interval in the partition of the variable support.

A location-scale model is considered to describe the distribution of the response conditionally on the covariates,

$$Y = \mu(\mathbf{z},\mathbf{x}) + \sigma(\mathbf{z},\mathbf{x})\varepsilon \ , \tag{1}$$

where $\mu(\mathbf{z},\mathbf{x})$ denotes the conditional location, $\sigma(\mathbf{z},\mathbf{x})$ the conditional dispersion, and ε an error term independent of $\mathbf{z}$ and $\mathbf{x}$ assumed to have fixed first and second order moments. One could for example assume that $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = 1$. The latter conditions lead to the interpretation of $\mu(\mathbf{z},\mathbf{x})$ and $\sigma(\mathbf{z},\mathbf{x})$ as the conditional mean and standard deviation, respectively. Other constraints are possible such as in Lambert (2013) where ε was assumed to have a zero median and a unit interquantile range, implying that $\mu(\mathbf{z},\mathbf{x})$ and $\sigma(\mathbf{z},\mathbf{x})$ had to be interpreted as the conditional median and interquantile range.

Assume that independent copies $(y_i, \mathbf{z}_i, \mathbf{x}_i)$ $(i = 1, \ldots, n)$ are observed on n units with the possibility of right or interval-censoring on y_i as described above. Additive models for the conditional location and dispersion of the response are specified,

$$\big(\mu(\mathbf{z}_i,\mathbf{x}_i)\big)_{i=1}^n = \left(\beta_0 + \sum_{k=1}^p \beta_k z_{ik} + \sum_{j=1}^J f_j^\mu(x_{ij})\right)_{i=1}^n = \mathbf{Z}\boldsymbol{\beta} + \sum_{j=1}^J \mathbf{f}_j^\mu \ , \tag{2}$$

$$\big(\log\sigma(\mathbf{z}_i,\mathbf{x}_i)\big)_{i=1}^n = \left(\delta_0 + \sum_{k=1}^p \delta_k z_{ik} + \sum_{j=1}^J f_j^\sigma(x_{ij})\right)_{i=1}^n = \mathbf{Z}\boldsymbol{\delta} + \sum_{j=1}^J \mathbf{f}_j^\sigma \ , \tag{3}$$

where $f_j^\mu(\cdot)$ and $f_j^\sigma(\cdot)$ denote smooth additive terms quantifying the effect of the jth quantitative covariate on the conditional mean and dispersion, $\mathbf{f}_j^\mu = \big(f_j^\mu(x_{ij})\big)_{i=1}^n$ and $\mathbf{f}_j^\sigma = \big(f_j^\sigma(x_{ij})\big)_{i=1}^n$ their values over units stacked in vectors, $\mathbf{Z}$ the $n \times (1+p)$ design matrix with a column of 1's for the intercept and one column per additional categorical covariate. Now consider a basis of $(L+1)$ cubic B-splines $\{s_{j\ell}^*(\cdot)\}_{\ell=1}^{L+1}$ associated to equally spaced knots on the range $(x_j^{\min}, x_j^{\max})$ of values for x_j. They are recentered for identification purposes in the additive model using $s_{j\ell}(\cdot) = s_{j\ell}^*(\cdot) - \frac{1}{x_j^{\max} - x_j^{\min}} \int_{x_j^{\min}}^{x_j^{\max}} s_{j\ell}^*(u)du$ $(\ell = 1, \ldots, L)$. Then, the additive terms in the conditional location and dispersion models can be approximated using linear combinations of these (recentered) B-splines, $\mathbf{f}_j^\mu = \left(\sum_{\ell=1}^L s_{j\ell}(x_{ij})\theta_{\ell j}^\mu\right)_{i=1}^n = \mathbf{S}_j\boldsymbol{\theta}_j^\mu$, $\mathbf{f}_j^\sigma = \left(\sum_{\ell=1}^L s_{j\ell}(x_{ij})\theta_{\ell j}^\sigma\right)_{i=1}^n = \mathbf{S}_j\boldsymbol{\theta}_j^\sigma$, where $[\mathbf{S}_j]_{i\ell} = s_{j\ell}(x_{ij})$, $\big(\boldsymbol{\theta}_j^\mu\big)_\ell = \theta_{\ell j}^\mu$ and $\big(\boldsymbol{\theta}_j^\sigma\big)_\ell = \theta_{\ell j}^\sigma$. Hence, using vectorial notations, the expressions for the conditional location and dispersion in (2) and (3) can be rewritten as $\big(\mu_i = \mu(\mathbf{z}_i,\mathbf{x}_i)\big)_{i=1}^n = \boldsymbol{\mathcal{X}}\boldsymbol{\psi}^\mu$, $\big(\sigma_i = \sigma(\mathbf{z}_i,\mathbf{x}_i)\big)_{i=1}^n = \exp\big(\boldsymbol{\mathcal{X}}\boldsymbol{\psi}^\sigma\big)$ with design matrix $\boldsymbol{\mathcal{X}} = [\mathbf{Z}, \mathbf{S}_1, \ldots, \mathbf{S}_J] = [\mathbf{Z}, \boldsymbol{\mathcal{S}}] \in \mathbb{R}^{n\times q}$; matrices of spline parameters (with one column per additive term) $\boldsymbol{\Theta}^\mu = [\boldsymbol{\theta}_1^\mu, \ldots, \boldsymbol{\theta}_J^\mu]$, $\boldsymbol{\Theta}^\sigma = [\boldsymbol{\theta}_1^\sigma, \ldots, \boldsymbol{\theta}_J^\sigma]$ in $\mathbb{R}^{L\times J}$; vectors of (stacked) regression parameters $\boldsymbol{\psi}^\mu = \big(\boldsymbol{\beta}, \mathrm{vec}(\boldsymbol{\Theta}^\mu)\big)$, $\boldsymbol{\psi}^\sigma = \big(\boldsymbol{\delta}, \mathrm{vec}(\boldsymbol{\Theta}^\sigma)\big)$ in $\mathbb{R}^q$, where $q = (1 + p + JL)$. Different subsets of covariates could be selected for the location and dispersion submodels. With p_1 (resp. p_2) covariates and a B-spline basis of size L_1 (resp. L_2) for each of the J_1 (resp. J_2) additive terms in the location (resp. dispersion) model, we would end up with design matrices $\boldsymbol{\mathcal{X}}^\mu = [\mathbf{Z}^\mu, \boldsymbol{\mathcal{S}}^\mu] \in \mathbb{R}^{n\times q_1}$ (resp. $\boldsymbol{\mathcal{X}}^\sigma = [\mathbf{Z}^\sigma, \boldsymbol{\mathcal{S}}^\sigma] \in \mathbb{R}^{n\times q_2}$) with $q_1 = (1 + p_1 + J_1L_1)$ (resp. $q_2 = (1 + p_2 + J_2L_2)$) such that $\big(\mu_i\big)_{i=1}^n = \boldsymbol{\mathcal{X}}^\mu\boldsymbol{\psi}^\mu$, $\big(\sigma_i\big)_{i=1}^n = \exp\big(\boldsymbol{\mathcal{X}}^\sigma\boldsymbol{\psi}^\sigma\big)$.

2.1. Penalized log-likelihood for the joint regression model

Estimation of the regression parameters and of the additive terms (for given penalty parameters) can be made using penalized likelihood. Denote by $\mathcal{D}$ the available data (including covariates) and by $f_\epsilon(\cdot\,;\boldsymbol{\phi})$ (resp. $S_\epsilon(\cdot\,;\boldsymbol{\phi})$) the conditional density (resp. survival function) of the standardized error term ϵ in (1) with a possible dependence on a set of parameters $\boldsymbol{\phi}$. The contribution $\ell_i=\ell_i(\boldsymbol{\psi}^\mu,\boldsymbol{\psi}^\sigma,\boldsymbol{\phi};\mathcal{D})$ of unit i to the log-likelihood $\ell(\boldsymbol{\psi}^\mu,\boldsymbol{\psi}^\sigma,\boldsymbol{\phi};\mathcal{D})=\sum_i \ell_i$ will depend on the censoring status of the observed response y_i:

- Uncensored $y_i=t_i$: then, the corresponding standardized error term e_i is equal to $r_i=(y_i-\mu_i)/\sigma_i$ with log-likelihood contribution $\ell_i=-\log\sigma_i+\log f_\epsilon(r_i)$.
- Right-censored at $y_i>t_i$: then, the corresponding standardized error term is $e_i>r_i=(t_i-\mu_i)/\sigma_i$ with log-likelihood contribution $\ell_i=\log S_\epsilon(r_i)$.
- Interval-censored with $y_i\in(y_i^L,y_i^R)$: then, the log-likelihood contribution is $\ell_i=\log\big(S(r_i^L)-S(r_i^R)\big)$ as $e_i\in(r_i^L,r_i^R)$ where $r_i^L=(y_i^L-\mu_i)/\sigma_i$ and $r_i^R=(y_i^R-\mu_i)/\sigma_i$.

Smoothness of the additive terms can be tuned by penalizing changes in differences of neighbour spline parameters (Eilers and Marx, 1996, 2010; Marx and Eilers, 1998). In a frequentist framework, this can be done by adding one penalty (to the log-likelihood) per additive term. When penalizing second-order differences in the location model, the penalty for the jth additive term ($j=1,\ldots,J_1$) becomes $\lambda_j^\mu\sum_{\ell=1}^{L_1-2}\{(\theta^\mu_{\ell+2,j}-\theta^\mu_{\ell+1,j})-(\theta^\mu_{\ell+1,j}-\theta^\mu_{\ell,j})\}^2=\lambda_j^\mu\sum_\ell\left(\mathbf{D}^\mu\boldsymbol{\theta}_j^\mu\right)_\ell^2=\boldsymbol{\theta}_j^{\mu\top}(\lambda_j^\mu\mathbf{P}^\mu)\boldsymbol{\theta}_j^\mu$, where $\mathbf{D}^\mu$ denotes the corresponding difference matrix and $\mathbf{P}^\mu=(\mathbf{D}^\mu)^\top\mathbf{D}^\mu$ the associated penalty matrix. At the limit, as $\lambda_j^\mu\to+\infty$, the estimated second-order differences will tend to zero, forcing the estimate of the function $f_j^\mu(x_j)$ to be linear. Similar penalties with penalty parameters λ_j^σ can be defined for each additive term in the dispersion model. Other penalty orders or penalty matrices could be preferred.

2.2. Bayesian specification

In a Bayesian framework, similar penalties arise through the specification of conditional priors for the spline parameters (Lang and Brezger, 2004), yielding for the jth additive terms in the location and dispersion models,

$$p(\boldsymbol{\theta}_j^\mu|\lambda_j^\mu)\propto\exp\left(-\frac{1}{2}\,\boldsymbol{\theta}_j^{\mu\top}(\lambda_j^\mu\mathbf{P}^\mu)\boldsymbol{\theta}_j^\mu\right)\quad;\quad p(\boldsymbol{\theta}_j^\sigma|\lambda_j^\sigma)\propto\exp\left(-\frac{1}{2}\,\boldsymbol{\theta}_j^{\sigma\top}(\lambda_j^\sigma\mathbf{P}^\sigma)\boldsymbol{\theta}_j^\sigma\right).$$

Assuming joint Normal priors for the intercepts and the regression parameters associated to the other covariates $\mathbf{z}$, $\boldsymbol{\beta}\sim\mathcal{N}\left(\tilde{\mathbf{b}},(\mathbf{Q}^\mu)^{-1}\right)$, $\boldsymbol{\delta}\sim\mathcal{N}\left(\tilde{\mathbf{d}},(\mathbf{Q}^\sigma)^{-1}\right)$, the joint priors for the regression and spline parameters in $\boldsymbol{\psi}^\mu$ and $\boldsymbol{\psi}^\sigma$ induce Gaussian Markov random fields (GMRF) (Rue and Held, 2005) as they can be written as

$$\begin{aligned}p(\boldsymbol{\psi}^\mu|\boldsymbol{\lambda}^\mu)&\propto\exp\left(-\frac{1}{2}\,(\boldsymbol{\psi}^\mu-\mathbf{b})^\top\mathbf{K}_\lambda^\mu(\boldsymbol{\psi}^\mu-\mathbf{b})\right)\;;\\ p(\boldsymbol{\psi}^\sigma|\boldsymbol{\lambda}^\sigma)&\propto\exp\left(-\frac{1}{2}\,(\boldsymbol{\psi}^\sigma-\mathbf{d})^\top\mathbf{K}_\lambda^\sigma(\boldsymbol{\psi}^\sigma-\mathbf{d})\right),\end{aligned}$$

where $\mathbf{b}=(\tilde{\mathbf{b}},\mathbf{0}_{J_1L_1})$, $\mathbf{K}_\lambda^\mu=\operatorname{diag}\big(\mathbf{Q}^\mu,\boldsymbol{\mathcal{P}}_\lambda^\mu\big)$, $\boldsymbol{\mathcal{P}}_\lambda^\mu=\boldsymbol{\Lambda}^\mu\otimes\mathbf{P}^\mu$, $[\boldsymbol{\Lambda}^\mu]_{jj'}=\delta_{jj'}\lambda_j^\mu$, $\mathbf{d}=(\tilde{\mathbf{d}},\mathbf{0}_{J_2L_2})$, $\mathbf{K}_\lambda^\sigma=\operatorname{diag}\big(\mathbf{Q}^\sigma,\boldsymbol{\mathcal{P}}_\lambda^\sigma\big)$, $\boldsymbol{\mathcal{P}}_\lambda^\sigma=\boldsymbol{\Lambda}^\sigma\otimes\mathbf{P}^\sigma$ and $[\boldsymbol{\Lambda}^\sigma]_{jj'}=\delta_{jj'}\lambda_j^\sigma$. Then the joint posterior for the parameters is

$$p(\boldsymbol{\psi}^\mu,\boldsymbol{\psi}^\sigma,\boldsymbol{\lambda}^\mu,\boldsymbol{\lambda}^\sigma,\boldsymbol{\phi}|\mathcal{D})\propto L(\boldsymbol{\psi}^\mu,\boldsymbol{\psi}^\sigma,\boldsymbol{\phi};\mathcal{D})\,p(\boldsymbol{\psi}^\mu|\boldsymbol{\lambda}^\mu)\;\,p(\boldsymbol{\psi}^\sigma|\boldsymbol{\lambda}^\sigma)\;\,p(\boldsymbol{\lambda}^\mu)\,p(\boldsymbol{\lambda}^\sigma)\,p(\boldsymbol{\phi}).$$

2.3. Estimation of $\boldsymbol{\psi}^\mu$ and $\boldsymbol{\psi}^\sigma$

The estimation of the regression parameters $\boldsymbol{\psi} = (\boldsymbol{\psi}^\mu, \boldsymbol{\psi}^\sigma) \in \mathbb{R}^{q_1+q_2}$ will be made iteratively and conditionally on the error density $f_\epsilon(\cdot\ ;\boldsymbol{\phi})$ and the penalty parameters $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^\mu, \boldsymbol{\lambda}^\sigma) \in \mathbb{R}_+^{J_1+J_2}$. It is based on their joint conditional posterior:

$$p(\boldsymbol{\psi}|\boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D}) \propto L(\boldsymbol{\psi}^\mu, \boldsymbol{\psi}^\sigma, \boldsymbol{\phi}; \mathcal{D})\ p(\boldsymbol{\psi}^\mu|\boldsymbol{\lambda}^\mu)\ p(\boldsymbol{\psi}^\sigma|\boldsymbol{\lambda}^\sigma). \tag{4}$$

The conditional posterior mode $\hat{\boldsymbol{\psi}}_\lambda$ for $\boldsymbol{\psi}$ given $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$ is computed using a Newton-Raphson (N-R) algorithm built upon the gradient $\mathbf{U}_\lambda$ and the Hessian $\mathbf{H}_\lambda$ of the log of (4),

$$\mathbf{U}_\lambda = \begin{pmatrix} \mathbf{U}_\lambda^\mu \\ \mathbf{U}_\lambda^\sigma \end{pmatrix} \quad ; \quad \mathbf{H}_\lambda = \begin{bmatrix} \mathbf{H}_\lambda^{\mu\mu} & \mathbf{H}_\lambda^{\mu\sigma} \\ \\ \mathbf{H}_\lambda^{\sigma\mu} & \mathbf{H}_\lambda^{\sigma\sigma} \end{bmatrix} \quad ,$$

with closed form expressions for these quantities given in Appendix A. It leads to Algorithm 1 for the estimation of the regression parameters $\boldsymbol{\psi}^\mu$ and $\boldsymbol{\psi}^\sigma$. At convergence, after a few iterations, one obtains the conditional posterior mode $\hat{\boldsymbol{\psi}}_\lambda$ with negative inverse Hessian $\Sigma_\lambda = (-\mathbf{H}_\lambda(\hat{\boldsymbol{\psi}}_\lambda))^{-1}$, yielding the following Laplace approximation to the conditional posterior of $\boldsymbol{\psi}$: $(\boldsymbol{\psi}|\boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D}) \dot\sim \mathcal{N}\left(\hat{\boldsymbol{\psi}}_\lambda, \Sigma_\lambda\right)$. Corrections to the so-obtained point estimates for the regression parameters could be made using the same type of arguments as with restricted maximum likelihood (REML) in the frequentist literature, see Jørgensen and Knudsen (2004) for some details in the framework of dispersion models. In our context, we explored the extraction of point estimates for $\boldsymbol{\psi}^\sigma$ using an approximation to its marginal posterior,

$$\begin{aligned} p(\boldsymbol{\psi}^\sigma|\boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D}) = \frac{p(\boldsymbol{\psi}^\mu, \boldsymbol{\psi}^\sigma|\boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D})}{p(\boldsymbol{\psi}^\mu|\boldsymbol{\psi}^\sigma, \boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D})} &= \frac{p(\hat{\boldsymbol{\psi}}_\lambda^\mu, \boldsymbol{\psi}^\sigma|\boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D})}{p(\hat{\boldsymbol{\psi}}_\lambda^\mu|\boldsymbol{\psi}^\sigma, \boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D})} \\ &\dot\propto L(\hat{\boldsymbol{\psi}}_\lambda^\mu, \boldsymbol{\psi}^\sigma, \boldsymbol{\phi}; \mathcal{D})\, p(\boldsymbol{\psi}^\sigma|\boldsymbol{\lambda}^\sigma)\, \left|-\mathbf{H}_\lambda^{\mu\mu}\right|^{-1/2}. \end{aligned}$$

It reduces the bias in the estimation of the intercept δ_0 with a growing impact when information gets scarce such as with small sample sizes or large censoring rates. However, we will not explore it further here, as our interest lies mainly in estimating the effects of covariates.

```
Algorithm 1: Estimation of $\boldsymbol{\psi}^\mu$ and $\boldsymbol{\psi}^\sigma$ (for given $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$)
Input: Penalty parameters $\boldsymbol{\lambda}$ and standardized error distribution
       parameters $\boldsymbol{\phi}$
Output: $\hat{\boldsymbol{\psi}}_\lambda$ (for given $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$) ; $\mathcal{D}_r$: the set of standardized residuals $r_i$
        (potentially right-censored or interval-censored $(r_i^L, r_i^R)$).
repeat
    1. Set $\mu_i \longleftarrow \mu(\mathbf{z}_i^\mu, \mathbf{x}_i^\mu; \boldsymbol{\psi}_\lambda^\mu)$ and $\sigma_i \longleftarrow \sigma(\mathbf{z}_i^\sigma, \mathbf{x}_i^\sigma; \boldsymbol{\psi}_\lambda^\sigma)$ for $1 \le i \le n$.
    2. Set $r_i \longleftarrow (y_i - \mu_i)/\sigma_i$ if $y_i$ observed or right-censored ; set
       $r_i^L \longleftarrow (y_i^L - \mu_i)/\sigma_i$ and $r_i^R \longleftarrow (y_i^R - \mu_i)/\sigma_i$ if $y_i$ interval-censored.
    3. Recompute vectors $\boldsymbol{\omega}^\mu$, $\boldsymbol{\omega}^\sigma$, $\mathbf{w}^{\mu\mu}$, $\mathbf{w}^{\sigma\sigma}$, $\mathbf{w}^{\mu\sigma}$ and set matrices
       $\mathbf{W}^{\mu\mu} \longleftarrow \mathrm{diag}(\mathbf{w}^{\mu\mu})$, $\mathbf{W}^{\sigma\sigma} \longleftarrow \mathrm{diag}(\mathbf{w}^{\sigma\sigma})$, $\mathbf{W}^{\mu\sigma} \longleftarrow \mathrm{diag}(\mathbf{w}^{\mu\sigma})$ using
       (A.3) and (A.4).
    4. Evaluate $\mathbf{U}_\lambda$ and $\mathbf{H}_\lambda$ using (A.1) and (A.2).
    5. Update $\boldsymbol{\psi}_\lambda \longleftarrow \boldsymbol{\psi}_\lambda - \mathbf{H}_\lambda^{-1}\mathbf{U}_\lambda$ with step-halving if found necessary
       from the monitoring of $p(\boldsymbol{\psi}|\boldsymbol{\lambda}, \boldsymbol{\phi}, D)$.
until $||\mathbf{U}_\lambda||_\infty < \epsilon$;
```

2.4. Selection of the penalty parameters $\boldsymbol{\lambda}^\mu$ and $\boldsymbol{\lambda}^\sigma$

Starting from the joint posterior for the model parameters, we have (with an implicit dependence on the standardized error distribution and its parameter(s) $\boldsymbol{\phi}$) the following identity for the marginal posterior of $\boldsymbol{\lambda}$:

$$p(\boldsymbol{\lambda}|\mathcal{D}) = \frac{p(\boldsymbol{\psi}, \boldsymbol{\lambda}|\mathcal{D})}{p(\boldsymbol{\psi}|\boldsymbol{\lambda}, \mathcal{D})}. \tag{5}$$

Given the conditional GMRF prior for $\boldsymbol{\psi}$, see Section 2.2, we conclude that the conditional posterior in the denominator is approximately Gaussian (Rue and Martino, 2009). Using a Laplace approximation, we obtain $(\boldsymbol{\psi}|\boldsymbol{\lambda}, \mathcal{D}) \mathrel{\dot\sim} \mathcal{N}\left(\hat{\boldsymbol{\psi}}_\lambda, \Sigma_\lambda\right)$, where $\hat{\boldsymbol{\psi}}_\lambda$ denotes the conditional posterior mode of $\boldsymbol{\psi}$ (obtained using Algorithm 1) and variance-covariance matrix

$$\Sigma_\lambda = \begin{bmatrix} -\mathbf{H}_\lambda^{\mu\mu} & -\mathbf{H}_\lambda^{\mu\sigma} \\ -\mathbf{H}_\lambda^{\sigma\mu} & -\mathbf{H}_\lambda^{\sigma\sigma} \end{bmatrix}^{-1} \quad , \tag{6}$$

with submatrix expressions given in (A.2), (A.3), (A.4), see also Tierney and Kadane (1986) for general arguments for such an approximation to the marginal posterior of $\boldsymbol{\lambda}$. Evaluating the RHS of (5) at $\hat{\boldsymbol{\psi}}_\lambda$ with the preceding Laplace approximation substituted in the denominator, we approximate $p(\boldsymbol{\lambda}|\mathcal{D})$ by

$$\tilde{p}(\boldsymbol{\lambda}|\mathcal{D}) \propto p(\hat{\boldsymbol{\psi}}_\lambda, \boldsymbol{\lambda}|\mathcal{D}) \left|\Sigma_\lambda^{-1}\right|^{-1/2}. \tag{7}$$

Wood and Fasiolo (2017, Section 4) obtained a similar starting expression to build their proposal for the selection of penalty parameters in an additive regression model with a parametric error distribution. In a full Bayesian approach, Gressani and Lambert (2018) also followed that strategy in cure survival models where splines were used to specify the baseline hazard function for susceptible subjects. It was further explored in the exponential family with generalized additive models in Gressani and Lambert (2021) and the associated R package `blapsr` available from the official R repository.

A direct maximization of (7) provides the desired point selection for $\boldsymbol{\lambda}$. We have also investigated an iterative two-step strategy alternating the update of $\boldsymbol{\lambda}^\mu$ and $\boldsymbol{\lambda}^\sigma$ to select a value denoted $\hat{\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda}$. Dropping the cross-derivatives between ψ^μ and ψ^σ in the expression for Σ_λ^{-1} in (6) yields the approximations

$$\begin{aligned} \tilde{p}(\boldsymbol{\lambda}^\mu|\boldsymbol{\lambda}^\sigma, \mathcal{D}) &\propto p(\hat{\boldsymbol{\psi}}_\lambda, \boldsymbol{\lambda}|\mathcal{D}) \left|\boldsymbol{\mathcal{X}}^{\mu\top}\mathbf{W}^{\mu\mu}\boldsymbol{\mathcal{X}}^\mu + \mathbf{K}_\lambda^\mu\right|^{-1/2} \quad ; \\ \tilde{p}(\boldsymbol{\lambda}^\sigma|\boldsymbol{\lambda}^\mu, \mathcal{D}) &\propto p(\hat{\boldsymbol{\psi}}_\lambda, \boldsymbol{\lambda}|\mathcal{D}) \left|\boldsymbol{\mathcal{X}}^{\sigma\top}\mathbf{W}^{\sigma\sigma}\boldsymbol{\mathcal{X}}^\sigma + \mathbf{K}_\lambda^\sigma\right|^{-1/2} \quad , \end{aligned} \tag{8}$$

with $\mathbf{W}^{\mu\mu}$ and $\mathbf{W}^{\sigma\sigma}$ given in Appendix A. Dropping the μ or σ superscript and letting

$$\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{S}}^\top\mathbf{W}\boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{S}}^\top\mathbf{W}\mathbf{Z}(\mathbf{Z}^\top\mathbf{W}\mathbf{Z} + \mathbf{Q})^{-1}\mathbf{Z}^\top\mathbf{W}\boldsymbol{\mathcal{S}}, \tag{9}$$

each determinant in (8) can be rewritten as

$$\left|\boldsymbol{\mathcal{X}}^\top\mathbf{W}\boldsymbol{\mathcal{X}} + \mathbf{K}_\lambda\right| = \left|\mathbf{Z}^\top\mathbf{W}\mathbf{Z} + \mathbf{Q}\right| \left|\boldsymbol{\mathcal{M}} + \boldsymbol{\mathcal{P}}_\lambda\right|,$$

where only the last factor directly depends on the penalty parameters $\boldsymbol{\lambda}$. Combined with (8) and taking $\lambda_j^\mu \sim \mathcal{G}\left(1, b^\mu = 10^{-4}\right)$, we conclude that

$$\begin{aligned} &\log\tilde{p}(\boldsymbol{\lambda}^\mu|\boldsymbol{\lambda}^\sigma, \mathcal{D}) \doteq \log p(\hat{\boldsymbol{\psi}}_\lambda, \boldsymbol{\lambda}|\mathcal{D}) - \frac{1}{2}\log\left|\boldsymbol{\mathcal{M}}^\mu + \boldsymbol{\mathcal{P}}_\lambda^\mu\right| \\ &= \ell(\hat{\boldsymbol{\psi}}_\lambda; \mathcal{D}) + \sum_{j=1}^{J_1}\left\{\frac{L_1 - \rho(\mathbf{P}^\mu)}{2}\log\lambda_j^\mu - \left(b^\mu + \frac{1}{2}(\hat{\boldsymbol{\theta}}_{j\lambda}^\mu)^\top\mathbf{P}^\mu\hat{\boldsymbol{\theta}}_{j\lambda}^\mu\right)\lambda_j^\mu\right\} \\ &\quad - \frac{1}{2}\log\left|\boldsymbol{\mathcal{M}}^\mu + \boldsymbol{\mathcal{P}}_\lambda^\mu\right|. \end{aligned} \tag{10}$$

The indirect dependence of the log-likelihood and of $\boldsymbol{\mathcal{M}}^\mu$ on λ^μ (through $\hat{\boldsymbol{\psi}}_\lambda$ and $\mathbf{W}^{\mu\mu}$) will be ignored during the computation of the gradient $\mathbf{U}^{\lambda^\mu}$ and Hessian $\mathbf{H}^{\lambda^\mu}$ as (non reported) numerical simulations suggest that this dependence is moderate. The simulation study in Section 3 also reports satisfactory estimates for the additive terms in the considered settings. Practically, in an iterative maximization of (10) using the N-R algorithm, we fix $\ell(\hat{\boldsymbol{\psi}}_\lambda;\mathcal{D})$ and $\boldsymbol{\mathcal{M}}^\mu$ at their values $\breve{\ell}$ and $\breve{\boldsymbol{\mathcal{M}}}^\mu$ at the beginning of the iteration, and compute the gradient and Hessian of

$$\begin{aligned}\log\breve{p}(\boldsymbol{\lambda}^\mu|\boldsymbol{\lambda}^\sigma,\mathcal{D}) = \breve{\ell} + \sum_{j=1}^{J_1}\left\{\frac{L_1-\rho(\mathbf{P}^\mu)}{2}\log\lambda_j^\mu - \left(b^\mu + \frac{1}{2}(\hat{\boldsymbol{\theta}}_{j\lambda}^\mu)^\top\mathbf{P}^\mu\hat{\boldsymbol{\theta}}_{j\lambda}^\mu\right)\lambda_j^\mu\right\} \\ -\frac{1}{2}\log\left|\breve{\boldsymbol{\mathcal{M}}}^\mu + \boldsymbol{\mathcal{P}}_\lambda^\mu\right| \ .\end{aligned}$$

Let $\breve{\mathcal{R}}_j^\mu = \breve{\mathcal{R}}_j^\mu(\boldsymbol{\lambda}^\mu) = \left(\breve{\boldsymbol{\mathcal{M}}}^\mu + \boldsymbol{\mathcal{P}}_\lambda^\mu\right)^{-1}\left((\mathbf{1}_j\mathbf{1}_j^\top)\otimes\mathbf{P}^\mu\right)$ for $j=1,\ldots,J_1$ where $\mathbf{1}_j$ denotes the jth unit vector. Then, using results on the derivative of determinants and after some algebra, on can show that

$$\begin{aligned}\left(\breve{\mathbf{U}}^{\lambda^\mu}\right)_j &= \frac{\partial\log\breve{p}(\boldsymbol{\lambda}^\mu|\boldsymbol{\lambda}^\sigma,\mathcal{D})}{\partial\lambda_j^\mu} \\ &= \frac{L_1-\rho(\mathbf{P}^\mu)}{2\lambda_j^\mu} - \left(b^\mu + \frac{1}{2}(\hat{\boldsymbol{\theta}}_{j\lambda}^\mu)^\top\mathbf{P}^\mu\hat{\boldsymbol{\theta}}_{j\lambda}^\mu\right) - \frac{1}{2}\mathrm{tr}\left(\breve{\mathcal{R}}_j^\mu\right), \\ -[\breve{\mathbf{H}}^{\lambda^\mu}]_{jk} &= -\frac{\partial^2\log\breve{p}(\boldsymbol{\lambda}^\mu|\boldsymbol{\lambda}^\sigma,\mathcal{D})}{\partial\lambda_j^\mu\partial\lambda_k^\mu} \\ &= \frac{L_1-\rho(\mathbf{P}^\mu)}{2(\lambda_j^\mu)^2}\delta_{jk} - \frac{1}{2}\mathrm{tr}\left(\breve{\mathcal{R}}_j^\mu\breve{\mathcal{R}}_k^\mu\right).\end{aligned}\tag{11}$$

Similar expressions can be obtained for $(\boldsymbol{\lambda}^\sigma|\boldsymbol{\lambda}^\mu,\mathcal{D})$ by switching the role of μ and σ as superscripts, giving $\breve{\mathbf{U}}^{\lambda^\sigma}$ and $\breve{\mathbf{H}}^{\lambda^\sigma}$. Define $\boldsymbol{\nu}^\mu\in\mathrm{I\!R}^{J_1}$ and $\boldsymbol{\nu}^\sigma\in\mathrm{I\!R}^{J_2}$ such that $\boldsymbol{\lambda}^\mu = \lambda_{\min}^\mu + \exp(\boldsymbol{\nu}^\mu)$ and $\boldsymbol{\lambda}^\sigma = \lambda_{\min}^\sigma + \exp(\boldsymbol{\nu}^\sigma)$ with $\lambda_{\min}^\mu$ and $\lambda_{\min}^\sigma$denoting the smallest desirable values for the penalty parameter of an additive term. Then, using the chain rule, one can show that the gradient and Hessian for functions

$$g(\boldsymbol{\nu}^\mu) = \log\tilde{p}(\boldsymbol{\lambda}^\mu|\boldsymbol{\lambda}^\sigma,\mathcal{D}) \ \ ; \ \ g(\boldsymbol{\nu}^\sigma) = \log\tilde{p}(\boldsymbol{\lambda}^\sigma|\boldsymbol{\lambda}^\mu,\mathcal{D}) \ , \tag{12}$$

are given by

$$\begin{aligned}(\breve{\mathbf{U}}^{\nu^\zeta})_j &= \exp(\nu_j^\zeta)(\breve{\mathbf{U}}^{\lambda^\zeta})_j \ , \\ (\breve{\mathbf{H}}^{\nu^\zeta})_{jk} &= \exp(\nu_j^\zeta+\nu_k^\zeta)(\breve{\mathbf{H}}^{\lambda^\zeta})_{jk} + \delta_{jk}\exp(\nu_j^\zeta)(\breve{\mathbf{U}}^{\lambda^\zeta})_j \ ,\end{aligned}\tag{13}$$

where $\zeta\in\{\mu,\sigma\}$, $1\le j,k\le J_1$ for $\boldsymbol{\nu}^\mu$ and $1\le j,k\le J_2$ for $\boldsymbol{\nu}^\sigma$. In the two-step strategy mentioned above and detailed in Algorithm 2, the penalty parameters are selected to maximize each function in (12) in an iterative procedure involving Newton-Raphson algorithms, yielding at convergence $\hat{\boldsymbol{\lambda}}^\mu$, $\hat{\boldsymbol{\lambda}}^\sigma$ and $\hat{\boldsymbol{\psi}}_\lambda$ for a given value of $\boldsymbol{\phi}$.

2.5. Nonparametric pivotal density

2.5.1. Density specification

Besides classical parametric choices for the distribution of the standardized error term ϵ, nonparametric forms could be preferred. Here, we propose to specify that distribution through the associated hazard $h_\epsilon(\cdot)$ function using a linear combination of K B-splines, $\log h_\epsilon(r) = \sum_{k=1}^K b_k(r)\phi_k$, where $\{b_k(\cdot) : k=1,\ldots,K\}$ denotes a

```
Algorithm 2: Selection of $\boldsymbol{\lambda}$ and estimation of $\boldsymbol{\psi}_\lambda$ (for given $\boldsymbol{\phi}$)
  Input: Spline parameters $\boldsymbol{\phi}$ specifying the standardized error density ;
         data $\mathcal{D}$ with a location-scale model specified in (2) and (3).
  Output: Selected $\boldsymbol{\lambda}$ and estimated $\hat{\boldsymbol{\psi}}_\lambda$ (for given $\boldsymbol{\phi}$) ; the resulting
          standardized residuals $\mathcal{D}_r$.
  repeat
      Compute $\hat{\boldsymbol{\psi}}_\lambda$ and $\mathcal{D}_r$ using Algorithm 1 for given $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$
      Set $\breve{\ell} \longleftarrow \ell(\hat{\boldsymbol{\psi}}_\lambda;\mathcal{D})$ using Sect. 2.1 and update $\breve{\boldsymbol{\mathcal{M}}}^\mu$, $\breve{\boldsymbol{\mathcal{M}}}^\sigma$ using (9)
      repeat
          Evaluate $\breve{\mathbf{U}}^{\nu^\mu}$ and $\breve{\mathbf{H}}^{\nu^\mu}$ using (11) & (13)
          $\boldsymbol{\nu}^\mu \longleftarrow \boldsymbol{\nu}^\mu - (\breve{\mathbf{H}}^{\nu^\mu})^{-1}\breve{\mathbf{U}}^{\nu^\mu}$
          $\boldsymbol{\lambda}^\mu \longleftarrow \lambda^\mu_{\min} + \exp(\boldsymbol{\nu}^\mu)$
      until $||\breve{\mathbf{U}}^{\nu^\mu}||_\infty < \epsilon$;
      repeat
          Evaluate $\breve{\mathbf{U}}^{\nu^\sigma}$ and $\breve{\mathbf{H}}^{\nu^\sigma}$ using (11) & (13)
          $\boldsymbol{\nu}^\sigma \longleftarrow \boldsymbol{\nu}^\sigma - (\breve{\mathbf{H}}^{\nu^\sigma})^{-1}\breve{\mathbf{U}}^{\nu^\sigma}$
          $\boldsymbol{\lambda}^\sigma \longleftarrow \lambda^\sigma_{\min} + \exp(\boldsymbol{\nu}^\sigma)$
      until $||\breve{\mathbf{U}}^{\nu^\sigma}||_\infty < \epsilon$;
  until convergence;
```

large B-spline basis associated to an equidistant grid of knots on the support of the distribution. Given the constraints $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = 1$, one can practically assume (using Chebyshev's theorem) that (most of) the probability mass is on compact support $(r_{\min}, r_{\max})$ $(= (-6, 6)$, say). Our approach is to some extent connected to the proposal made by Cai et al. (2002) with a (truncated) linear spline basis in a mixed model framework. We go further here by considering interval-censored data and moment constraints for the underlying density function. Note that starting from the hazard function to estimate the underlying distribution does not imply that the variable must be positive. The only requirement is the selection of a (conservative) lower bound for the support of the standardized error term. A spline approximation to the log-density could also be considered (Eilers and Marx, 1996; Kooperberg and Stone, 1991; Lambert and Eilers, 2009; Lambert, 2011), but a construct based on the hazard function turns out to be analytically more convenient to handle censored data, see below.

2.5.2. Density estimation from i.i.d. right-censored data

We now detail how we propose to estimate the spline coefficients $\boldsymbol{\phi}$ in the framework of Bayesian P-splines from potentially right- or even interval-censored data gathered in $\mathcal{D}_r$.

Denote by $\{\mathcal{J}_j = [a_{j-1}, a_j)\}_{j=1}^J$ a partition of $(r_{\min}, r_{\max})$ into a very large number J of bins of equal width Δ with midpoints $\{u_j\}_{j=1}^J$. Given a random sample of n i.i.d. observations r_i $(i = 1, \ldots, n)$ for a potentially right-censored (coded by $d_i = 0$ and 1 otherwise) variable ε, let $k_j = \sum_{i=1}^n k_{ij}$ and $n_j = \sum_{i=1}^n n_{ij}$ with $k_{ij} = \mathbb{1}(r_i \in \mathcal{J}_j)\,\mathbb{1}(d_i = 1)$ and $n_{ij} = \mathbb{1}(r_i \geq a_{j-1}) = \mathbb{1}(r_i \in \cup_{s\geq j}\mathcal{J}_s)$. The log-likelihood for the estimation of the spline parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$ from right-censored data can be written as

$$\ell(\boldsymbol{\phi}|\mathcal{D}_r) = \sum_{i=1}^n \left\{d_i \log h_\epsilon(r_i) - H_\epsilon(r_i)\right\} \approx \sum_{j=1}^J (k_j \log h_j - n_j h_j \Delta) \ , \qquad (14)$$

with $h_j = h_\epsilon(u_j) = \exp\{\sum_{k=1}^K b_k(u_j)\phi_k\}$ where the approximation in (14) comes from data binning and quadrature to approximate the cumulative hazard function. Following Eilers and Marx (1996), we penalize rth order differences of successive spline parameters, yielding the penalized log-likelihood, $\ell_p(\boldsymbol{\phi}|\tau,\mathcal{D}_r) = \ell(\boldsymbol{\phi}|\mathcal{D}_r) - \frac{\tau}{2}\boldsymbol{\phi}^\top\mathbf{P}\boldsymbol{\phi}$, with penalty matrix $\mathbf{P}$ of rank $(K-r)$. Given the expressions for the gradient and Hessian,

$$\mathbf{U}_\tau(\boldsymbol{\phi}) = \frac{\partial \ell_p}{\partial \boldsymbol{\phi}} = \mathbf{B}^\top(\mathbf{k} - \mathbf{nh}\Delta) - \tau\mathbf{P}\boldsymbol{\phi} \ ; \tag{15}$$

$$-\mathbf{H}_\tau(\boldsymbol{\phi}) = -\frac{\partial^2 \ell_p}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^\top} = \mathbf{B}^\top\mathrm{diag}(\mathbf{nh}\Delta)\mathbf{B} + \tau\mathbf{P}, \tag{16}$$

where $[\mathbf{B}]_{jk} = b_k(u_j)$, $\mathbf{k} = (k_j)_{j=1}^J$, $\mathbf{n} = (n_j)_{j=1}^J$, $\mathbf{h} = (h_j)_{j=1}^J$, one can use the (fast converging) Newton-Raphson procedure to obtain spline parameter estimates for a given value of the penalty parameter τ, with at each iteration, $\boldsymbol{\phi} \longleftarrow \boldsymbol{\phi} - \big(\mathbf{H}_\tau(\boldsymbol{\phi})\big)^{-1}\mathbf{U}_\tau(\boldsymbol{\phi})$, yielding $\hat{\boldsymbol{\phi}}_\tau$ at convergence.

2.5.3. Inclusion of interval-censored data

The contribution of interval-censored units to k_j and n_j can also be included and reevaluated at every iteration of the preceding Newton-Raphson procedure. Denote the hazard and density estimates from the previous iteration by $\tilde{h}_\epsilon(\cdot)$ and $\tilde{f}_\epsilon(\cdot) = \tilde{h}_\epsilon(\cdot)\exp(-\tilde{H}_\epsilon(\cdot))$, and let $\tilde{\pi}_j = \int_{\mathcal{J}_j}\tilde{f}_\epsilon(r)dr \approx \tilde{f}_\epsilon(u_j)\Delta$. Consider an interval-censored observation $r_i \in (r_i^L, r_i^R)$ and let $\mathcal{G}_i = \{j : \mathcal{J}_j \cap (r_i^L, r_i^R) \neq \emptyset\}$. Then, the contribution of unit i to the previously defined k_j and n_j are given by $k_{ij} = \tilde{\pi}_j / \sum_{s\in\mathcal{G}_i}\tilde{\pi}_s \ \mathbb{1}(j\in\mathcal{G}_i)$ and $n_{ij} = \mathbb{1}(j < \min\mathcal{G}_i) + \sum_{s=j}^{\max\mathcal{G}_i}\tilde{\pi}_s / \sum_{s\in\mathcal{G}_i}\tilde{\pi}_s\mathbb{1}(j\in\mathcal{G}_i)$, respectively. At convergence, the procedure in Section 2.5.2 with, now, interval-censored data entering the computation of k_j and n_j will provide an estimate $\hat{\boldsymbol{\phi}}_\tau$ of the spline parameters $\boldsymbol{\phi}$ for given τ and, hence, of the density estimate underlying the potentially right- or interval-censored observations.

2.5.4. Density estimation with moment constraints

Constraints on the mean and variance of the underlying distribution can also be enforced. More generally, consider a set of (potentially) nonlinear constraints $F_s(\boldsymbol{\phi}) = f_s$ $(s = 1,\ldots,S)$ shortly denoted vectorially by $\boldsymbol{F}(\boldsymbol{\phi}) = \mathbf{f}$. At every iteration of the preceding Newton-Raphson procedure, we suggest to linearize each constraint using a first-order Taylor expansion about the current estimate $\tilde{\boldsymbol{\phi}}$ of the spline parameters, $\tilde{F}_s(\boldsymbol{\phi}) = F_s(\tilde{\boldsymbol{\phi}}) + \tilde{\mathbf{v}}_s^\top(\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}})$ with $\tilde{\mathbf{v}}_s = \frac{\partial F_s(\tilde{\boldsymbol{\phi}})}{\partial\boldsymbol{\phi}}$. Hence, letting $\tilde{\mathbf{V}} = [\tilde{\mathbf{v}}_1,\ldots,\tilde{\mathbf{v}}_S]^\top \in \mathbb{R}^{S\times K}$, a linearized version of the constraints is $\tilde{\mathbf{V}}\boldsymbol{\phi} = \tilde{\mathbf{c}}$ with $\tilde{\mathbf{c}} = \tilde{\mathbf{V}}\tilde{\boldsymbol{\phi}} + (\mathbf{f} - \boldsymbol{F}(\tilde{\boldsymbol{\phi}}))$. The estimation of the spline parameters under these linearized constraints can be made using the Lagrangian

$$G(\boldsymbol{\phi},\boldsymbol{\omega}) = \ell_p(\boldsymbol{\phi}|\tau,\mathcal{D}_r) - \boldsymbol{\omega}^\top(\tilde{\mathbf{V}}\boldsymbol{\phi} - \tilde{\mathbf{c}}), \tag{17}$$

with Lagrange multipliers $\boldsymbol{\omega}$. Practically, at every iteration of a Newton-Raphson procedure, the preceding values $(\tilde{\boldsymbol{\phi}},\tilde{\boldsymbol{\omega}})$ of the spline parameters and Lagrange multipliers are updated using

$$\begin{pmatrix}\tilde{\boldsymbol{\phi}}\\ \tilde{\boldsymbol{\omega}}\end{pmatrix} \longleftarrow \begin{pmatrix}\tilde{\boldsymbol{\phi}}\\ \tilde{\boldsymbol{\omega}}\end{pmatrix} - \begin{bmatrix}\frac{\partial^2\ell_p(\tilde{\boldsymbol{\phi}}|\tau,\mathcal{D}_r)}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^\top} & -\tilde{\mathbf{V}}^\top \\ -\tilde{\mathbf{V}} & \mathbf{0}\end{bmatrix}^{-1} \begin{pmatrix}\frac{\partial\ell_p(\tilde{\boldsymbol{\phi}}|\tau,\mathcal{D}_r)}{\partial\boldsymbol{\phi}} - \tilde{\mathbf{V}}^\top\tilde{\boldsymbol{\omega}} \\ -\tilde{\mathbf{V}}\tilde{\boldsymbol{\phi}} + \tilde{\mathbf{c}}\end{pmatrix}, \tag{18}$$

with partial derivatives of the penalized log-likelihood given in (15) and (16).

Now consider specific constraints on the spline parameters based on the first two moments ($S = 2$) of the density, remembering that $f(u_j) = h_j \exp(-H_j)$ (and letting $\Delta \to 0^+$):

$$\mathbb{E}(\epsilon) = \mu_\epsilon = 0 \Leftrightarrow F_1(\boldsymbol{\phi}) = \sum_{j=1}^{J} u_j h_j \exp(-H_j)\Delta = 0 = f_1 \ ;$$

$$\mathbb{V}(\epsilon) = \sigma_\epsilon^2 = 1 \Leftrightarrow F_2(\boldsymbol{\phi}) = \sum_{j=1}^{J} u_j^2 h_j \exp(-H_j)\Delta - F_1(\boldsymbol{\phi})^2 = 1 = f_2 \ .$$

Let $\tilde{h}_j = \tilde{h}_\epsilon(u_j)$, $\tilde{H}_j = \sum_{\ell \le j} \tilde{h}_j \Delta$, $\tilde{f}_j = \tilde{h}_j \exp(-\tilde{H}_j)$ and $b_{jk} = b_k(u_j)$. Then, one can show that $\tilde{\mathbf{V}}_{1k} = \frac{\partial F_1(\tilde{\boldsymbol{\phi}})}{\partial \phi_k} = \sum_{j=1}^{J} u_j \tilde{f}_j \Delta \left(b_{jk} - \sum_{\ell \le j} b_{\ell k} \tilde{h}_\ell \Delta \right)$ and $\tilde{\mathbf{V}}_{2k} = \frac{\partial F_2(\tilde{\boldsymbol{\phi}})}{\partial \phi_k} = \sum_{j=1}^{J} u_j^2 \tilde{f}_j \Delta \left(b_{jk} - \sum_{\ell \le j} b_{\ell k} \tilde{h}_\ell \Delta \right) - 2F_1(\tilde{\boldsymbol{\phi}})\, \tilde{\mathbf{V}}_{1k}$. Combining these last results with the elements from Sections 2.5.2 and 2.5.3, one can estimate the spline parameters underlying the hazard and, hence, the density, for given (potentially) right- or interval-censored data and penalty parameter τ. The following section is devoted to the selection of τ.

2.5.5. Selection of the penalty parameter τ

Given the following priors,

$$\tau \sim \mathcal{G}(1, b) \quad ; \quad p(\boldsymbol{\phi}|\tau) \propto \tau^{\frac{K-r}{2}} \exp\left(-\frac{\tau}{2} \boldsymbol{\phi}^\top \mathbf{P} \boldsymbol{\phi} \right), \tag{19}$$

the joint posterior for the spline and the penalty parameters $(\boldsymbol{\phi}, \tau)$ are

$$p(\boldsymbol{\phi}, \tau | D) \propto \exp\{\ell(\boldsymbol{\phi}|\mathcal{D}_r)\}\, p(\boldsymbol{\phi}|\tau)\, p(\tau) = \exp\{\ell_p(\boldsymbol{\phi}|\tau, \mathcal{D}_r)\}\, \tau^{\frac{K-r}{2}}\, p(\tau). \tag{20}$$

Using the same arguments as in Section 2.4 for $(\boldsymbol{\psi}|\boldsymbol{\lambda}, D)$, the conditional posterior for the spline parameters, $p(\boldsymbol{\phi}|\tau, D) \propto \exp\{\ell_p(\boldsymbol{\phi}|\tau, \mathcal{D}_r)\}$, can be shown to be approximately

$$(\boldsymbol{\phi}|\tau, \mathcal{D}_r) \mathrel{\dot\sim} \mathcal{N}\left(\hat{\boldsymbol{\phi}}_\tau, \hat{\Sigma}_\tau \right), \tag{21}$$

where $\hat{\boldsymbol{\phi}}_\tau$ denotes the conditional posterior mode (equal to the penalized MLE of $\boldsymbol{\phi}$ given τ, see Sections 2.5.2 and 2.5.3), $\hat{\Sigma}_\tau^{-1} = -\mathbf{H}_\tau(\hat{\boldsymbol{\phi}}_\tau) = \mathbf{B}^\top \mathbf{W}_\tau \mathbf{B} + \tau \mathbf{P}$, see (16), with $\mathbf{W}_\tau = \mathrm{diag}(\mathbf{w}_\tau)$, $\mathbf{w}_\tau = \mathbf{n}\hat{\mathbf{h}}_\tau \Delta$ and $\hat{\mathbf{h}}_\tau$ giving the estimated hazard at the bin midpoints when $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}_\tau$. Given that the number of observations $(\mathbf{k})_j$ in bin $\mathcal{J}_j$ has expected value $(\mathbf{w})_j = (\mathbf{n}\mathbf{h}\Delta)_j$, one might reasonably approximate the last variance-covariance matrix by $\hat{\Sigma}_\tau^{-1} \approx \mathbf{B}^\top \mathbf{W} \mathbf{B} + \tau \mathbf{P}$ with $\mathbf{W} = \mathrm{diag}(\mathbf{k})$, thereby restricting its explicit dependence on τ to the $\tau \mathbf{P}$ term. The marginal posterior for τ is given by

$$p(\tau|\mathcal{D}_r) = \frac{p(\boldsymbol{\phi}, \tau|\mathcal{D}_r)}{p(\boldsymbol{\phi}|\tau, \mathcal{D}_r)} \mathrel{\dot\propto} p(\hat{\boldsymbol{\phi}}_\tau, \tau|\mathcal{D}_r)\, |\mathbf{B}^\top \mathbf{W} \mathbf{B} + \tau \mathbf{P}|^{-1/2} \ , \tag{22}$$

with the approximation coming from (21) and the substitution of $\mathbf{W}_\tau$ by $\mathbf{W}$. Now consider a singular value decomposition of penalty matrix, $\mathbf{P} = \mathbf{U}\boldsymbol{\Upsilon}\mathbf{U}^\top$, where $\mathbf{U} = [\mathbf{U}_1 \ \mathbf{U}_0]$, $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_K$, $\boldsymbol{\Upsilon} = \mathrm{blockdiag}(\boldsymbol{\Upsilon}_1, \mathbf{0}_r)$, with the last r diagonal elements of $\boldsymbol{\Upsilon} = \mathrm{diag}(\boldsymbol{\upsilon})$ being zero. Then, using properties of determinants and defining $\tilde{\mathbf{B}} = \mathbf{W}^{1/2}\mathbf{B}\mathbf{U}$, $\tilde{\mathbf{B}}_1 = \mathbf{W}^{1/2}\mathbf{B}\mathbf{U}_1$, $\tilde{\mathbf{B}}_0 = \mathbf{W}^{1/2}\mathbf{B}\mathbf{U}_0$, $\mathbf{M} = \tilde{\mathbf{B}}_1^\top \tilde{\mathbf{B}}_1 - \tilde{\mathbf{B}}_1^\top \tilde{\mathbf{B}}_0 (\tilde{\mathbf{B}}_0^\top \tilde{\mathbf{B}}_0)^{-1} \tilde{\mathbf{B}}_0^\top \tilde{\mathbf{B}}_1$, one has

$$|\mathbf{B}^\top \mathbf{W} \mathbf{B} + \tau \mathbf{P}| = |\tilde{\mathbf{B}}_0^\top \tilde{\mathbf{B}}_0|\, |\boldsymbol{\Upsilon}_1|\, \tau^{K-r} \prod_{j=1}^{K-r} \left(1 + \frac{n\tilde{m}_j}{\tau} \right) \ , \tag{23}$$

where $\widetilde{\mathbf{M}} = \frac{1}{n}\boldsymbol{\Upsilon}_1^{-1/2}\mathbf{M}\boldsymbol{\Upsilon}_1^{-1/2}$ has eigenvalues $\{\tilde{m}_j\}_{j=1}^{K-r}$ independent of τ. Combining (19), (20), (22) and (23), one has

$$\begin{aligned}\log p(\tau|\mathcal{D}_r) &\doteq \ell_p(\hat{\boldsymbol{\phi}}_\tau|\tau,\mathcal{D}_r) + \log p(\tau) - \frac{1}{2}\sum_{j=1}^{K-r}\log\left(1+\frac{n\tilde{m}_j}{\tau}\right)\\ &= \ell(\hat{\boldsymbol{\phi}}_\tau|\mathcal{D}_r) - \tau\left(b + \frac{1}{2}\hat{\boldsymbol{\phi}}_\tau^\top\mathbf{P}\hat{\boldsymbol{\phi}}_\tau\right) - \frac{1}{2}\sum_{j=1}^{K-r}\log\left(1+\frac{n\tilde{m}_j}{\tau}\right), \qquad (24)\end{aligned}$$

suggesting Algorithm 3 to select τ.

For example, with a dataset of size $n = 1\,000$ including 40% uncensored, 40% interval-censored and 20% right-censored data, the selection of τ and the estimation of $K = 50$ B-spline parameters (an unnecessary very large K used to challenge Algorithm 3) took 6 iterations and one tenth of a second using pure R code on a low-end desktop computer. The specified constraints on the values of the first two moments were perfectly met (up to the numerical tolerance specified by the user) in this example and when estimating the standardized error distribution in the location-scale models fitted to the many datasets generated in Section 3 under different right- and interval-censoring schemes.

2.6. Algorithm for fitting the NP additive location-scale model

We now have all the necessary ingredients for fitting the nonparametric double additive location-scale model (NP-DALSM) from possibly right- or even interval-censored data. Algorithm 4 is iterative and alternates the estimation of the regression and spline parameters in the location and dispersion submodels (Step 1) with the estimation of the standardized error density (Step 2). Possible starting values for that algorithm are obtained by:

- Assuming a Gaussian standardized error distribution ;
- Discarding right-censored data and setting interval-censored ones to their mid-point value, yielding a reduced response vector $\tilde{\boldsymbol{y}}$ with an associated design matrix $\tilde{\mathcal{X}}^\mu$ for the additive location submodel ;
- Setting the elements in penalty vectors $\boldsymbol{\lambda}^\mu$ and $\boldsymbol{\lambda}^\sigma$ to a moderately large value (100, say) ;
- Estimating $\boldsymbol{\psi}^\mu$ using penalized LS: $\boldsymbol{\psi}^\mu \longleftarrow \left(\tilde{\mathcal{X}}^{\mu\top}\tilde{\mathcal{X}}^\mu + \mathbf{K}^\mu_\lambda\right)^{-1}\tilde{\mathcal{X}}^{\mu\top}\tilde{\boldsymbol{y}}$;
- Fixing $\boldsymbol{\psi}^\sigma$ to zero, except its first component δ_0 set to the log of the mean squared error.

One major advantage of our proposal is the simultaneous update and estimation of the regression and spline parameters in the additive submodels for location and dispersion. The penalty parameters tuning the smoothness of the additive terms are jointly and automatically selected using a Newton-Raphson procedure based on approximate analytical expressions for the gradient and Hessian of their marginal posterior. And last but not least, the standardized error distribution is also estimated through the underlying (log-)hazard expressed as a linear combination of (penalized) P-splines with a penalty parameter selected to maximize its marginal posterior density. The whole procedure is able to handle right- or interval-censored response data.

Convergence of Algorithm 4 is very fast with the suggested initial conditions. It is implemented using pure R code in the package `DALSM` that can be obtained from the author's website or from the GitHub repository at `https://github.com/plambertULiege/`.

Algorithm 3: Constrained density estimation (selection of τ and estimation of $\boldsymbol{\phi}$ under constraints $\boldsymbol{F}(\boldsymbol{\phi}) = \mathbf{f}$)

```
Goal: Smooth estimation of a density distribution $f(e)$ on a compact
      support $(e_{\min}, e_{\max})$ and modelled as
      $f(e|\boldsymbol{\phi}) = h(e|\boldsymbol{\phi}) \exp(-\int_{e_{\min}}^{e} h(s|\boldsymbol{\phi}) ds)$ where
      $h(e|\boldsymbol{\phi}) = \exp\left(\sum_k b_k(e)\phi_k\right)$.
Input: Set of independent data $\mathcal{D}_r$ potentially right- or interval-censored ;
       the distribution support ; constraints $\boldsymbol{F}(\boldsymbol{\phi}) = \mathbf{f}$.
Output: Penalized estimation $\hat{\boldsymbol{\phi}}_\tau$ of $\boldsymbol{\phi}$ following the selection of the
        penalty parameter $\tau$.

Principle:
repeat
    $\hat{\boldsymbol{\phi}}_\tau \longleftarrow \mathrm{argmax}_\phi\, p(\boldsymbol{\phi}|\tau, \mathcal{D}_r)$ under the constraints $\boldsymbol{F}(\boldsymbol{\phi}) = \mathbf{f}$ ;
    $\hat{\boldsymbol{\phi}} \longleftarrow \hat{\boldsymbol{\phi}}_\tau$
    $\tau \longleftarrow \mathrm{argmax}_\tau\ p(\hat{\boldsymbol{\phi}}, \tau|\mathcal{D}_r)\, |\mathbf{B}^\top \mathbf{W}\mathbf{B} + \tau\mathbf{P}|^{-1/2}$.
until convergence;

Practically:
repeat
    1. Given the current estimate for $\tau$, maximize the Lagrangian in (17) by
       repeating the Newton-Raphson step in (18) till convergence to $\hat{\boldsymbol{\phi}}_\tau$.
    2. Update $\tau$ by using the fixed-point method on the partial derivative of
       (24) w.r.t. $\tau$ set to zero. Specifically:
    (a) Set $\hat{\boldsymbol{\phi}} \longleftarrow \hat{\boldsymbol{\phi}}_\tau$
    (b) repeat
            $$\tau \longleftarrow \sum_{j=1}^{K-r} \frac{n\tilde{m}_j}{\tau + n\tilde{m}_j} \Big/ \left(2b + \boldsymbol{\phi}^\top \mathbf{P}\boldsymbol{\phi}\right).$$
        until convergence;
until convergence;
At convergence, it yields $(\tau, \hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}_\tau)$ and $\hat{h}(\cdot) = \exp\left(\sum_{k=1}^{K} b_k(\cdot)\hat{\phi}_k\right)$.
```

Algorithm 4: Fitting the nonparametric DALSM model (NP-DALSM)

```
Goal: Fit of the NP-DALSM model described in (1), (2), (3) given data $\mathcal{D}$.
Input: Data $\mathcal{D}$ and DALSM model specification.
Output: Selected $\boldsymbol{\lambda}$, estimated regression parameters $\boldsymbol{\psi}_\lambda$ and standardized
        error density $f_\epsilon(\cdot|\boldsymbol{\phi})$.

repeat
    1. Select $\boldsymbol{\lambda}$, estimate $\boldsymbol{\psi}_\lambda$ and compute $\mathcal{D}_r$ using Algorithm 2 with data
       $\mathcal{D}$ for a given $\boldsymbol{\phi}$ ;
    2. Update $\boldsymbol{\phi}$ using Algorithm 3 with data $\mathcal{D}_r$ and constraints
       $F_1(\boldsymbol{\phi}) = \mathbb{E}(\epsilon|\boldsymbol{\phi}) = 0$ and $F_2(\boldsymbol{\phi}) = \mathbb{V}(\epsilon|\boldsymbol{\phi}) = 1$.
until convergence;
```

3. Simulation study

An extended simulation study was made to evaluate the performances of the proposed algorithm to fit the nonparametric additive location-scale model. The data were simulated with conditional location and dispersion given by, respectively,

$$\begin{aligned} \mu(\mathbf{z}^{\mu}, \mathbf{x}^{\mu}) &= (\beta_0 + \beta_1 z_1^{\mu} + \beta_2 z_2^{\mu}) + f_1^{\mu}(x_1^{\mu}) + f_2^{\mu}(x_2^{\mu}), \\ \log \sigma(\mathbf{z}^{\sigma}, \mathbf{x}^{\sigma}) &= (\delta_0 + \delta_1 z_1^{\sigma} + \beta_2 z_2^{\sigma}) + f_1^{\sigma}(x_1^{\sigma}) + f_2^{\sigma}(x_2^{\sigma}). \end{aligned} \tag{25}$$

Different combinations of sample sizes n $(= 1500, 500, 250)$, right-censoring $(RC = 0\%, 25\%, 50\%)$ rates and interval-censoring $(IC = 0\%, 25\%, 50\%)$ rates were considered. The standardized error term (with mean 0 and variance 1) in (1) was taken to have a Normal mixture distribution, $\epsilon \sim .8\,\mathcal{N}\left(-0.414, 0.538^2\right) + .2\,\mathcal{N}\left(1.655, 0.646^2\right)$, see Fig. B.8. For each of the n units, the pair of covariates $(p_1 = p_2 = 2)$ with linear effects in (25) were independently generated from Bernoulli and Normal distributions, $z_1^{\mu}, z_1^{\sigma} \sim \text{Bern}(.6)$; $z_2^{\mu}, z_2^{\sigma} \sim \mathcal{N}(0,1)$, with regression parameters $\boldsymbol{\beta} = (1.6, .3, .75)$, $\boldsymbol{\delta} = (-.5, -.03, .01)$. Two $(= J_1 = J_2)$ additive terms per regression submodel were added, $f_1^{\mu}(x) = .113 - .4\sqrt{x}\sin(1.2\pi x)$, $f_2^{\mu}(x) = .586 - .3(x^2 + .3)^{-1}$, $f_1^{\sigma}(x) = -0.158 + 0.15x + 0.25x^2$, $f_2^{\sigma}(x) = 12(x - 0.5)^3$, with $x_1^{\mu}, x_2^{\mu}, x_1^{\sigma}, x_1^{\sigma}$ generated independently from a uniform distribution on $(0, 1)$, see the solid curves on Fig. B.6 for a graphical representation. For each of the n units, covariates were first sampled to define the underlying first and second order (conditional) moments in (25), yielding μ_i and σ_i for the ith unit. The associated uncensored response was then obtained using $y_i = \mu_i + \sigma_i e_i$ with e_i sampled from the Normal mixture. Right-censoring was created randomly and independently of the underlying response and covariates using an exponential distribution $C_i \sim \text{Exp}(\lambda)$ with λ selected to reach the desired percentage RC of right-censored responses. The observed response was then defined as $t_i = \min\{y_i, c_i\}$ with *observation* indicator $\delta_i = I(c_i > y_i)$. The non right-censored data (for which $\delta_i = 1$) were subsequently interval-censored with probability $IC/(1 - RC)$ with, then, y_i only reported to lie in (y_i^L, y_i^R) where $y_i^L = y_i - 1.5u_i\sigma(Y)$ and $y_i^R = y_i + 1.5(1 - u_i)\sigma(Y)$ with $u_i \sim U_{(0,1)}$, yielding an interval of width equal to 1.5 the marginal standard deviation of the response.

The double additive location-scale model (DALSM) was fitted by assuming a nonparametric (NP) or a Normal ($\mathcal{N}$) density for the error term. Under the working Normality hypothesis, the sandwich estimator (White, 1982) was preferred over the model-based one for the variance-covariance of the regression and spline parameter estimates. A report on the detailed simulation results can be found in Appendix B. In summary, our simulation study suggests that the estimation of covariate effects on location and dispersion can be made with negligible biases even when the sample size is small (as compared to the number of parameters in the model), and whatever the nonparametric or Normal assumption made on the error term. But potentially large efficiency gains can be made under the NP assumption as compared to an approach with a working normality hypothesis. These gains tend to decrease with the amount of right-censoring as it inevitably affects the accuracy of the estimation of the right tail of the error density in the NP approach. The error density is properly estimated in the absence of right-censoring even with a rather small sample size and a large interval-censoring rate. But the combination of a small n and a large right-censoring rate somehow decrease the quality of the expected reconstruction as the available information on the right tail of the error distribution becomes sparse and incomplete. Then, the smallest component in the Normal mixture tends to be flattened around its mode. Right-censoring also impacts the estimation of the intercept δ_0 in the dispersion sub-model with a negative bias growing with the right-censoring rate whatever the assumption made on the error distribution.

The simulation study was repeated with data generated using the same submodels for the location and dispersion parts, but with a normal distribution for the error term. The detailed simulation results (available upon request) indicate that the estimates obtained by fitting the NP-DALSM or $\mathcal{N}$-DALSM models are very similar with comparable biases and efficiencies reported for the estimation of covariate effects in the fixed and additive parts, even with a small sample size ($n = 250$) or with the largest censoring rates considered. Given the potentially large efficiency gains in the non-normal case, this suggests that the NP-DALSM model is a recommendable option for estimating a double additive location-scale model when prior information on the error distribution is limited.

4. Application

The proposed application involves interval- and right-censored responses. The data of interest come from the European Social Survey (European Social Survey Round 8 Data, 2016). We focus on the money available per person in Belgian households for respondents aged 25-55 when the main source of income comes from wages or salaries ($n = 756$). Each person reports the total net monthly income of the household in one of 10 decile-based intervals: 1: $< 1\,120$ ($n_1 = 8$), 2: $[1\,120, 1\,399]$ ($n_2 = 13$), 3: $[1\,400, 1\,719]$ ($n_3 = 47$), 4: $[1\,720, 2\,099]$ ($n_4 = 53$), 5: $[2\,100, 2\,519]$ ($n_5 = 82$), 6: $[2\,520, 3\,059]$ ($n_6 = 121$), 7: $[3\,060,\ 3\,739]$ ($n_7 = 167$), 8: $[3\,740, 4\,529]$ ($n_8 = 126$), 9: $[4\,530, 5\,579]$ ($n_9 = 74$), 10: $\geq 5\,580$ euros ($n_{10} = 65$).

We model the relationship between disposable income per person (91.4% interval-censored, 8.6% right-censored) and the availability of (at least) two income (64.2%) in the household, as well as the age (`Age`: 41.0 ± 8.83 years) and number of years of education completed (`Educ`: 14.9 ± 3.34 years) by the respondent. This individualized income is obtained by dividing the household one by the OECD-modified equivalence scale (Hagenaars et al., 1994), as recommended by the Statistical Office of the European Union (EUROSTAT). The first adult in the household contributes to 1.0 to that scale, each person aged at least 14 adds .5 to it, while each younger member brings an extra .3 to the household weight. For example, a respondent aged 31 declaring a household net monthly income in the interval $(3\,060, 3\,740)$ euros with a partner aged 34 and 4 children aged 15, 10, 9 and 3 would be associated to an OECD-modified scale of 2.9 and an interval-censored response of $(1\,055.2, 1\,289.7)$ euros (available per person).

The nonparametric double additive location-scale model (NP-DALSM) described in Section 2 with the flexible error density from Section 2.5 was fitted using Algorithm 4: 10 ($=L$) and 20 ($=K$) B-splines were taken to model the additive terms and the log hazard of the error distribution, respectively. The response was rescaled in thousand euros. The algorithm converged after 5 iterations in less than one second using the author's R package `DALSM` on a low-end desktop computer. Parameter es-

Fixed	Location			Dispersion		
effects	$\hat\beta$	s.e.	CI 95%	$\hat\delta$	s.e.	CI 95%
`Intercept`	1.572	0.062	(1.450, 1.695)	-0.436	0.086	(-0.604, -0.268)
`TwoIncomes`	0.252	0.051	(0.152, 0.352)	-0.042	0.070	(-0.179, 0.094)

Additive	Location		Dispersion	
terms	e.d.f.	CI 95%	e.d.f.	CI 95%
`Age`	4.2	(2.3, 5.0)	3.3	(1.5, 4.4)
`Educ`	3.6	(1.8, 4.7)	3.3	(1.5, 4.3)

Table 1: Belgian income data (ESS 2016): fixed effect estimates and effective degrees of freedom (e.d.f.) (with 95% credible intervals) for the additive terms in the NP-DALSM model with the income response in thousand euros.

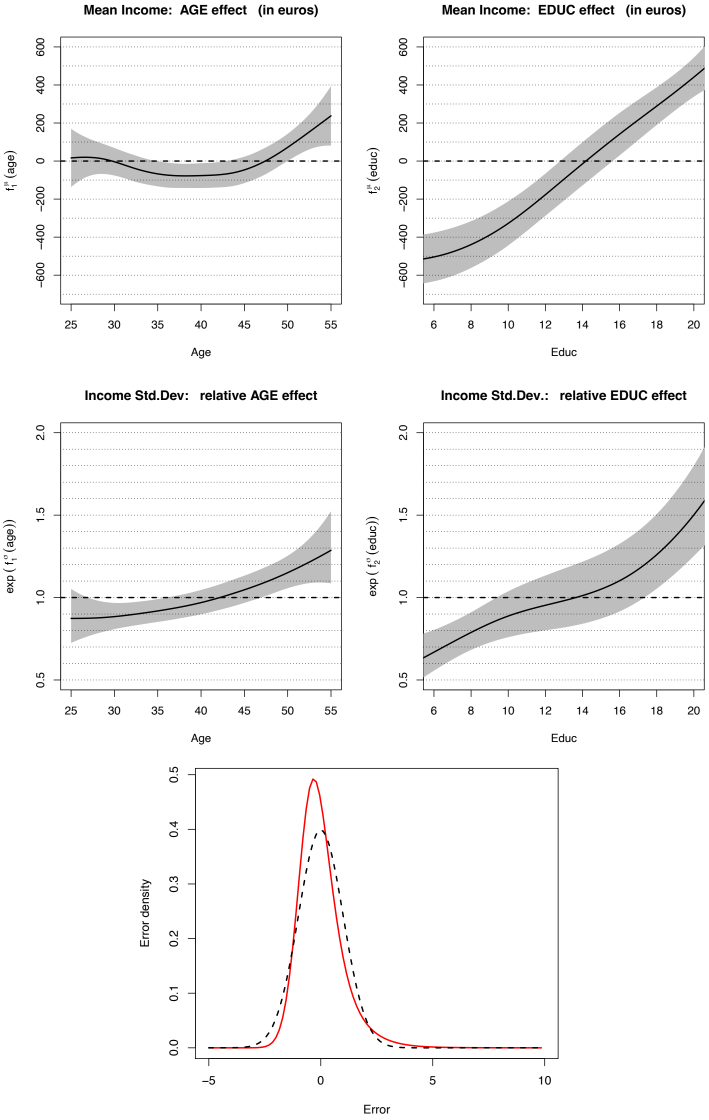


Figure 1: Belgian income data (ESS 2016): estimated additive terms in the NP-DALSM model with pointwise 95% credible intervals ; Row 1 (effects on location): $f_1^\mu(\text{Age})$ and $f_2^\mu(\text{Educ})$ rescaled in euros ; Row 2 (relative effects on dispersion): $\exp\left(f_1^\sigma(\text{Age})\right)$ and $\exp\left(f_2^\sigma(\text{Educ})\right)$; Row 3: Estimated error density (solid line) compared to the standard Normal (dashed line).

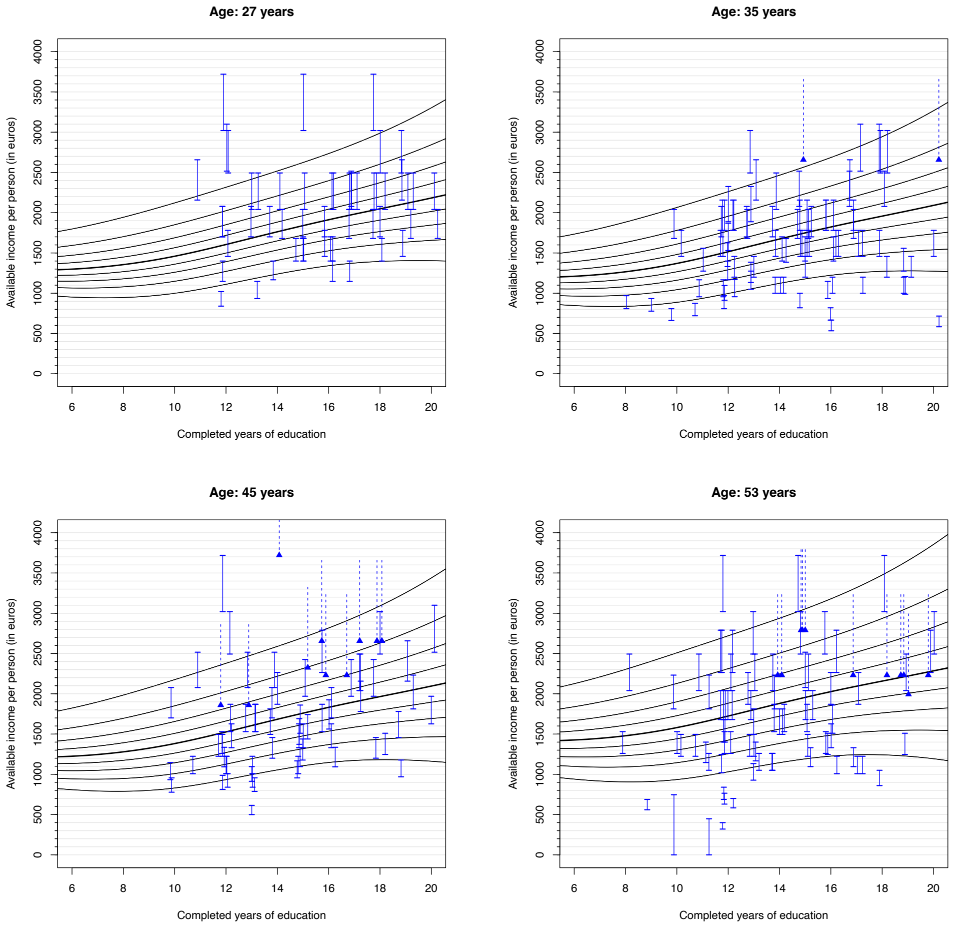


Figure 2: Belgian income data (ESS 2016): fitted conditional deciles (rescaled in euros) for the income per person in two-income households.

timates quantifying the effect of the TwoIncomes binary indicator on the conditional mean and the log of the standard deviation can be found in Table 1, suggesting an average increase of 252 euros when the respondent and his/her partner are in paid work (conditionally on Age and Educ), while the effect on dispersion is not statistically significant. The effects of Age and Educ on the conditional mean and dispersion can be visualized on the first and second rows of Fig. 1, respectively, with the corresponding estimated additive terms. The amount of money available per person in the household tends to decrease with age (see $f_1^\mu(\mathrm{Age})$) between approximately 27 and 37 (most likely due the arrival of children in the family) and to increase after 40 (probably thanks to wage increase with seniority and the departure of children). The dispersion, reported as the exponential of the additive term, $\exp(f_1^\sigma(\mathrm{Age}))$, significantly increases with Age with an acceleration over 30. However, the dominant effect comes from the respondent's level of education, with a difference of about 1 000 euros (in expected disposable income per person) between a less educated (6 years) and a highly educated (20 years) respondent, see $f_2^\mu(\mathrm{Educ})$. The effect on dispersion is also large, see $\exp(f_2^\sigma(\mathrm{Educ}))$, with essentially an important contrast between less and highly educated respondents, the latter group showing the largest heterogeneity. Indeed, while most low-skilled people have difficulty finding employment or are confined to low-paying occupations, a university degree offers a wide variety of opportunities ranging from a moderately-paid government job to an executive position in a multinational company in the chemical, pharmaceutical or financial sectors. The estimated density for the error term can also be seen at the bottom of Fig. 1, with a right-skewed shape clearly distinguishable from the Gaussian one often explicitly or implicitly assumed when fitting location-scale regression models. The resulting estimates for the deciles of the income available per person for varying education levels and ages are pictured on Fig. 2. Interval- and right-censored data are represented as intervals and dashed semi-intervals, respectively (with horizontal noise added to untie respondents sharing the same age). The combined nonlinear impacts of age and education discussed earlier on the distribution of disposable income per person are now clearly visible.

5. Discussion

The proposed nonparametric double additive location-scale model (NP-DALSM) is a fast and efficient alternative to parametric location-scale models. Unlike moment-based estimation approaches such as the generalized method of moments (see e.g. Wang et al., 2014), it provides a full estimation of the conditional distribution of the response, that can be used to understand and visualize how it is qualitatively and quantitatively affected by covariates. The density of the error distribution is estimated from possibly right- or interval-censored responses under moment constraints. The penalty parameters controlling the smoothness of the additive terms in the location and dispersion submodels are automatically selected using approximations to their marginal posteriors. These are obtained by substituting Laplace approximations to the conditional posterior of the spline parameters, see Section 2.4. Simulations suggest that the effects of covariates are properly estimated with negligible biases in the estimation of regression parameters and additive terms. However, biases during the estimation of the intercept in the dispersion submodel may appear when right-censoring rates are high, while some additive terms may be excessively smoothed (as they should be) when information becomes scarce. For example, it may result from a combination of large right-censoring rates and small sample sizes (relative to the large number of parameters to be estimated). The simulation study also suggests that the uncertainties in the estimates are correctly quantified as the effective coverages of the credible intervals for the parameters measuring the effects of covariates on location and dispersion are consistent with their nominal values.

The nonparametric specification with P-splines of (the log-hazard function underlying) the error density can markedly increase the efficiency of regression parameter and additive term estimates over results under a working Normality hypothesis, while reducing the risk of misleading conclusions following from a misspecified non-normal parametric error density. While our proposal extends to nonparametric errors and interval-censored settings some aspects of the remarkable work by Wood and Fasiolo (2017) or Wood (2017), several issues still need to be studied in that specific framework. Model validation is one topic, with the presence of interval-censored data complicating the capacity to diagnose misspecification from partially observed residuals. Model selection should also be investigated. Obvious starting solutions would consist in computing information criteria such as AIC and BIC with the number of parameters replaced by effective dimensions (Komárek et al., 2005). Uncertainty in the selection of penalty parameters can also be taken into account, see Wood et al. (2016), Wood (2017, Section 6.11) or Gressani and Lambert (2021) for additional perspectives. More elaborate procedures for testing the necessity to include an additive term (in location or dispersion) or to opt for a simpler linear form could be developed in our framework. From a Bayesian perspective (Rossel and Rubio, 2019), they should be built using a combination of the conditional posterior for the spline parameters entering the additive term of interest and the marginal posterior for the associated penalty parameter. Nonlinear and smooth interactions between covariates could also be added to the location and dispersion parts in the same way as Lee and Durbán (2011) and Rodríguez-Álvarez et al. (2018) with the conditional mean in mixed models.

Acknowledgments

The author would like to thank the Referees and the Associate Editor for their very helpful comments and suggestions. He also acknowledges the support of the ARC project IMAL (grant 20/25-107) financed by the Wallonia-Brussels Federation and granted by the Académie universitaire Louvain.

Appendix A. Closed form expressions for $\mathbf{U}_{\boldsymbol{\lambda}}$ and $\mathbf{H}_{\boldsymbol{\lambda}}$

Conditionally on $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^\mu, \boldsymbol{\lambda}^\sigma)$ and $\boldsymbol{\phi}$, one can obtain the following closed form expressions for the gradient $\mathbf{U}_\lambda$ and Hessian $\mathbf{H}_\lambda$ of the log posterior of the regression parameters $\boldsymbol{\psi} = (\boldsymbol{\psi}^\mu, \boldsymbol{\psi}^\sigma)$ in (4):

$$\begin{aligned}
\mathbf{U}^\mu_\lambda &= \mathbf{U}^\mu(\boldsymbol{\psi}^\mu|\boldsymbol{\lambda}) = \frac{\partial \log p(\boldsymbol{\psi}|\boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D})}{\partial \boldsymbol{\psi}^\mu} = \boldsymbol{\mathcal{X}}^{\mu\top}\boldsymbol{\omega}^\mu - \mathbf{K}^\mu_\lambda \boldsymbol{\psi}^\mu, \\
\mathbf{U}^\sigma_\lambda &= \mathbf{U}^\sigma(\boldsymbol{\psi}^\sigma|\boldsymbol{\lambda}) = \frac{\partial \log p(\boldsymbol{\psi}|\boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D})}{\partial \boldsymbol{\psi}^\sigma} = \boldsymbol{\mathcal{X}}^{\sigma\top}\boldsymbol{\omega}^\sigma - \mathbf{K}^\sigma_\lambda \boldsymbol{\psi}^\sigma,
\end{aligned} \tag{A.1}$$

and

$$\begin{aligned}
\mathbf{H}^{\mu\mu}_\lambda &= \mathbf{H}^{\mu\mu}(\boldsymbol{\psi}|\boldsymbol{\lambda}) = \frac{\partial^2 \log p(\boldsymbol{\psi}|\boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D})}{\partial \boldsymbol{\psi}^\mu \partial \boldsymbol{\psi}^{\mu\top}} = -\left(\boldsymbol{\mathcal{X}}^{\mu\top}\mathbf{W}^{\mu\mu}\boldsymbol{\mathcal{X}}^\mu + \mathbf{K}^\mu_\lambda\right) , \\
\mathbf{H}^{\sigma\sigma}_\lambda &= \mathbf{H}^{\sigma\sigma}(\boldsymbol{\psi}|\boldsymbol{\lambda}) = \frac{\partial^2 \log p(\boldsymbol{\psi}|\boldsymbol{\lambda}, \boldsymbol{\phi}, \mathcal{D})}{\partial \boldsymbol{\psi}^\sigma \partial \boldsymbol{\psi}^{\sigma\top}} = -\left(\boldsymbol{\mathcal{X}}^{\sigma\top}\mathbf{W}^{\sigma\sigma}\boldsymbol{\mathcal{X}}^\sigma + \mathbf{K}^\sigma_\lambda\right) , \\
\mathbf{H}^{\mu\sigma}_\lambda &= \mathbf{H}^{\mu\sigma}(\boldsymbol{\psi}|\boldsymbol{\lambda}) = \frac{\partial^2 \log p(\boldsymbol{\psi}|\boldsymbol{\lambda}^\sigma, \boldsymbol{\psi}^\sigma, \boldsymbol{\phi}, \mathcal{D})}{\partial \boldsymbol{\psi}^\mu \partial \boldsymbol{\psi}^{\sigma\top}} = -\left(\boldsymbol{\mathcal{X}}^{\mu\top}\mathbf{W}^{\mu\sigma}\boldsymbol{\mathcal{X}}^\sigma\right) ,
\end{aligned} \tag{A.2}$$

with vectors $\boldsymbol{\omega}^\mu$,$\boldsymbol{\omega}^\sigma$ in $\mathbb{R}^n$ and diagonal matrices $\mathbf{W}^{\mu\mu} = \mathrm{diag}(\mathbf{w}^{\mu\mu})$, $\mathbf{W}^{\sigma\sigma} = \mathrm{diag}(\mathbf{w}^{\sigma\sigma})$, $\mathbf{W}^{\mu\sigma} = \mathrm{diag}(\mathbf{w}^{\mu\sigma})$ in $\mathbb{R}^{n\times n}$ defined below. Rewriting the error density as $f_\epsilon(\cdot) = h_\epsilon(\cdot)\exp[-H_\epsilon(\cdot)]$ where $H_\epsilon(\cdot) = -\log S_\epsilon(\cdot)$ and $h_\epsilon(\cdot) = f_\epsilon(\cdot)/S_\epsilon(\cdot)$, we obtain the following expressions (depending on the censoring status of the response) for the aforementioned vectors and matrices:

Uncensored or right-censored t_i : if d_i is the observation indicator, then

$$
\begin{aligned}
\boldsymbol{\omega}_i^\mu &= -\frac{1}{\sigma_i}\left(d_i\frac{h_i'}{h_i} - h_i\right) \; ; \; \boldsymbol{\omega}_i^\sigma = -d_i r_i \frac{h_i'}{h_i} - d_i + r_i h_i \; , \\
\mathbf{w}_i^{\mu\mu} &= \frac{1}{\sigma_i^2}\left[d_i\left\{\left(\frac{h_i'}{h_i}\right)^2 - \frac{h_i''}{h_i}\right\} + h_i'\right] , \\
\mathbf{w}_i^{\mu\sigma} &= \frac{1}{\sigma_i}\left[d_i\left\{\left(\frac{h_i'}{h_i}\right)^2 r_i - \frac{h_i'}{h_i} - \frac{h_i''}{h_i} r_i\right\} + h_i' r_i + h_i\right] \; , \\
\mathbf{w}_i^{\sigma\sigma} &= d_i\left\{\left(\frac{h_i'}{h_i}\right)^2 r_i^2 - \frac{h_i'}{h_i} r_i - \frac{h_i''}{h_i} r_i^2\right\} + h_i' r_i^2 + h_i r_i \\
&= \sigma_i r_i \mathbf{w}_i^{\mu\sigma} \; ,
\end{aligned}
\tag{A.3}
$$

where $r_i = (t_i - \mu_i)/\sigma_i$, $h_i = h_\epsilon(r_i)$, $h_i' = \frac{dh_\epsilon(r_i)}{dr}$, $h_i'' = \frac{d^2 h_\epsilon(r_i)}{dr^2}$;

Interval-censored with $y_i \in (y_i^L, y_i^R)$:

$$
\begin{aligned}
\boldsymbol{\omega}_i^\mu &= \frac{1}{\sigma_i}\frac{f_\epsilon(r_i^L) - f_\epsilon(r_i^R)}{S_\epsilon(r_i^L) - S_\epsilon(r_i^R)} \; ; \; \boldsymbol{\omega}_i^\sigma = \frac{r_i^L f_\epsilon(r_i^L) - r_i^R f_\epsilon(r_i^R)}{S_\epsilon(r_i^L) - S_\epsilon(r_i^R)} \; , \\
\mathbf{w}_i^{\mu\mu} &= \frac{1}{\sigma_i^2}\left[\frac{f_\epsilon(r_i^L)g(r_i^L) - f_\epsilon(r_i^R)g(r_i^R)}{S_\epsilon(r_i^L) - S_\epsilon(r_i^R)} + \left\{\frac{f_\epsilon(r_i^L) - f_\epsilon(r_i^R)}{S_\epsilon(r_i^L) - S_\epsilon(r_i^R)}\right\}^2\right] , \\
&= \frac{1}{\sigma_i^2}\frac{f_\epsilon(r_i^L)g(r_i^L) - f_\epsilon(r_i^R)g(r_i^R)}{S_\epsilon(r_i^L) - S_\epsilon(r_i^R)} + (\boldsymbol{\omega}_i^\mu)^2 \; , \\
\mathbf{w}_i^{\sigma\sigma} &= \frac{r_i^L f_\epsilon(r_i^L)m(r_i^L) - r_i^R f_\epsilon(r_i^R)m(r_i^R)}{S_\epsilon(r_i^L) - S_\epsilon(r_i^R)} + \left\{\frac{r_i^L f_\epsilon(r_i^L) - r_i^R f_\epsilon(r_i^R)}{S_\epsilon(r_i^L) - S_\epsilon(r_i^R)}\right\}^2 \\
&= \frac{r_i^L f_\epsilon(r_i^L)m(r_i^L) - r_i^R f_\epsilon(r_i^R)m(r_i^R)}{S_\epsilon(r_i^L) - S_\epsilon(r_i^R)} + (\boldsymbol{\omega}_i^\sigma)^2 \; , \\
\mathbf{w}_i^{\mu\sigma} &= \frac{1}{\sigma_i}\frac{f_\epsilon(r_i^L)m(r_i^L) - f_\epsilon(r_i^R)m(r_i^R)}{S_\epsilon(r_i^L) - S_\epsilon(r_i^R)} + \boldsymbol{\omega}_i^\mu \boldsymbol{\omega}_i^\sigma \; ,
\end{aligned}
\tag{A.4}
$$

where $r_i^L = \frac{(y_i^L - \mu_i)}{\sigma_i}$, $r_i^R = \frac{(y_i^R - \mu_i)}{\sigma_i}$, $g(r) = \frac{h_\epsilon'(r)}{h_\epsilon(r)} - h_\epsilon(r)$ and $m(r) = 1 + rg(r)$.

Appendix B. Detailed simulation results

The double additive location-scale model (DALSM) was fitted by assuming a nonparametric (NP) or a Normal ($\mathcal{N}$) density for the error term with 10 (=L) B-splines (associated to equidistant knots on the range of each covariate) to reconstruct each of the additive terms and 20 (=K) B-splines (associated to equidistant knots on $(-6, 6)$) to estimate the (log of the hazard function underlying the) nonnormal standardized error density. Figures B.3, B.4 and B.5 report on the estimation of the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ for each of the three sample sizes for the nine possible combinations of right- and interval-censoring rates. The boxplots inform us on the (sampling) distribution of the parameter estimates (in grey for NP and white for $\mathcal{N}$) over the $S = 500$ replicates, R.E. indicates the Relative Efficiency (defined as the ratio of the mean squared errors) under a working normality hypothesis (a value smaller than 1.0 suggesting than the NP approach is preferable), while E.C. reports the Effective Coverage of 95% credible intervals (computed as $\hat{\theta} \pm 1.96$ s.e.(θ)). Whatever the considered sample size and the (NP or normal) assumption made on the standardized error distribution, the quantification of covariate effects on location and dispersion is made with negligible biases, except for the intercept δ_0 in the dispersion part where negative biases arise under right-censoring and tend to increase with the RC rate. In addition, whatever the sample size and the amount of censoring, mean squared errors for estimators of covariate effects are (nearly) always smaller when the error distribution is estimated nonparametrically, suggesting potentially important efficiency gains under the NP model. These gains tend to decrease when information gets sparse with decreasing sample sizes or increasing (interval- or right-) censoring rates. Except for δ_0 when a bias arises, the effective coverages of credible intervals are close to their nominal value 95% whatever the assumption made on the error distribution, suggesting that the standard errors were properly quantified and that the sampling distributions of the selected parameter estimators are close to normality.

Report on the estimation of the additive terms can be found in Tables B.2, B.3 and B.4. Whatever the sample size and censoring rates, the absolute biases averaged over the covariate support $(0, 1)$ are very small, at the exception of $f_2^\sigma(x)$ for values of x close to zero or one when the sample size gets small and the right-censoring rate is large. Then, given the sparse information available, the estimate of this additive term tends to be oversmoothed. It probably explains part of the bias reported during the estimation of the intercept δ_0. This is illustrated in Fig. B.6 and B.7 when the interval-censoring rate is 0% or 50%, respectively, for increasing right-censoring rates. The wider dark grey envelope (connecting successive intervals containing 95% of the additive term estimates $f_j^\mu(x)$ or $f_j^\sigma(x)$ over the S replicates) also indicate that the working Normality hypothesis for the error term yields less efficient estimates than under the NP assumption (with light-grey envelopes). This is confirmed numerically by the relative efficiency values reported in the preceding tables. The effective coverages of 95% credible intervals for $f_j^\mu(x)$ or $f_j^\sigma(x)$ averaged over the support $(0, 1)$ of the covariate and the S replicates are close to their nominal values whatever the sample size and the amount of censoring, except for $f_2^\sigma(x)$ under heavy right-censoring as, then, the effective coverage can be moderately smaller than expected.

The estimates of the NP error density (averaged over the S replicates) are given in Fig. B.8 for different combinations of right- and interval-censoring rates. Whatever the sample size and in the absence of right-censoring, the density is very well estimated with an excellent performance of the selection procedure for the underlying smoothness parameter (cf. Section 2.5.5). Large right-censoring rates tend to have an important negative effect on the quality of the reconstruction as it reduces the ability to detect or to correctly estimate the position of the second mode of the standardized error density. Combined with a large interval-censoring rate and a small sample size, it can even result in a right-skewed unimodal average density estimate (see the dotted curve at the bottom right of the figure) with the smallest component in the Normal mixture tending to be flattened around its mode. It contributes to the decreasing (but still existing) efficiency gains of the NP approach (over the working normal hypothesis for the error term) in these settings.

References

Breiman, L. and J. H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association 80*(391), 580–598.

Cai, T., R. J. Hyndman, and M. P. Wand (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics 11*(4), 784–798.

Croux, C., I. Gijbels, and I. Prosdocimi (2012). Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics 68*(1), 31–44.

Eilers, P. H. and B. D. Marx (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics 2*(6), 637–653.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*, 89–102.

European Social Survey Round 8 Data (2016). Data file edition 2.1. NSD - Norwegian Centre for Research Data, Norway.

Fan, J. and I. Gijbels (1994). Censored regression: Local linear approximations and their applications. *Journal of the American Statistical Association 89*(426), 560–570.

Gijbels, I. and I. Prosdocimi (2012). Flexible mean and dispersion function estimation in extended generalized additive models. *Communications in Statistics - Theory and Methods 41*(16-17), 3259–3277.

Gressani, O. and P. Lambert (2018). Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. *Computational Statistics and Data Analysis 124*, 151–167.

Gressani, O. and P. Lambert (2021). Laplace approximation for fast Bayesian inference in generalized additive models based on P-splines. *Computational Statistics and Data Analysis 124*. doi:10.1016/j.csda.2020.107088

Hagenaars, A., K. De Vos, and A. Zaidi (1994). *Poverty statistics in the late 1980's: research based on micro-data*. Luxembourg: Office for Official Publications of the European Communities.

Hastie, T. and R. Tibshirani (1986). Generalized additive models. *Statistical Science 1*(3), 297–318.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. London: Chapman & Hall.

Heuchenne, C. and I. Van Keilegom (2010). Estimation in nonparametric location-scale regression models with censored data. *Annals of the Institute of Statistical Mathematics 62*(3), 439–463.

Jørgensen, B. and S. J. Knudsen (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics 31*(1), 93–114.

Komárek, A., E. Lesaffre, and J. F. Hilton (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics 14*(3), 726–745.

Kooperberg, C. and C. J. Stone (1991). A study of logspline density estimation. *Computational Statistics and Data Analysis 12*(3), 327–347.

Lambert, P. (2011). Smooth semiparametric and nonparametric Bayesian estimation of bivariate densities from bivariate histogram data. *Computational Statistics and Data Analysis 55*(1), 429–445.

Lambert, P. (2013). Nonparametric additive location-scale models for interval censored data. *Statistics and Computing 23*, 75–90.

Lambert, P. and P. H. Eilers (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics and Data Analysis 53*(4), 1388–1399.

Lambert, P. and J. K. Lindsey (1999). Analysing financial returns by using regression models based on non-symmetric stable distributions. *Journal of the Royal Statistical Society. Series C: Applied Statistics 48*(3), 409–424.

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

Lee, D. J. and M. Durbán (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling 11*(1), 49–69.

Lee, Y., J. Nelder, and Y. Pawitan (2006). *Generalized Additive Models with Random Effects: Unified Analysis via H-likelihood*. Boca Raton: Chapman & Hall / CRC.

Marx, B. D. and P. H. C. Eilers (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis 28*(2), 193–209.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A 135*, 370–384.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Applied Statistics 54*(3), 507–554.

Rodríguez-Álvarez, M. X., M. P. Boer, F. A. van Eeuwijk, and P. H. Eilers (2018). Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics 23*, 52–71.

Rossell, D. and F. J. Rubio (2019). Additive Bayesian variable selection under censoring and misspecification. *arXiv preprint*, arXiv:1907.13563.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications (Monographs on Statistics and Applied Probability)*. Chapman & Hall/CRC.

Rue, H. and S. Martino (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B 71*(2), 319–392.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 1986*(393), 82–86.

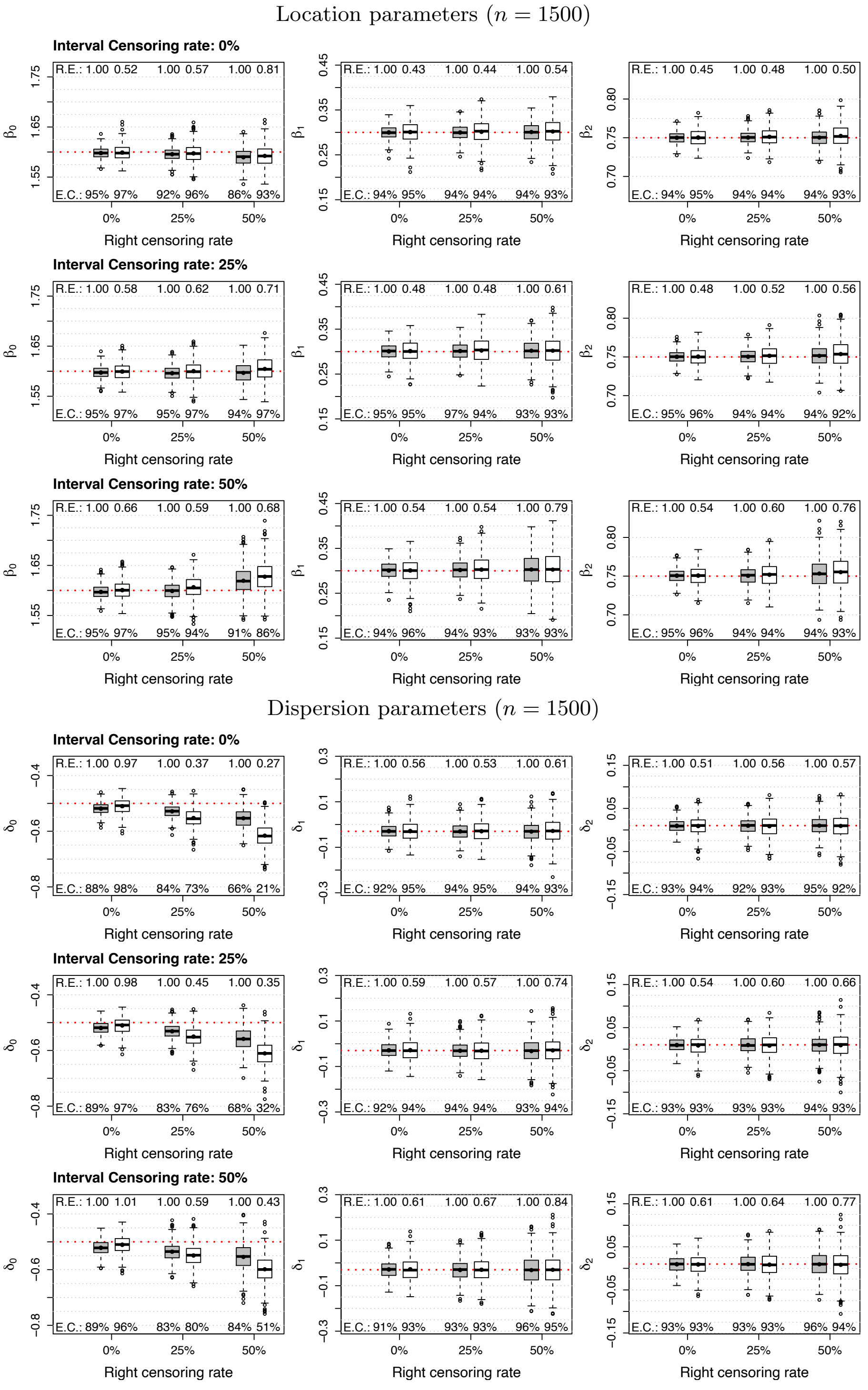
Wang, L., L. Xue, A. Qu, and H. Liang (2014). Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *Annals of Statistics 42*(2), 592–624.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R (2nd Edition)*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Wood, S. N. and M. Fasiolo (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics 73*(4), 1071–1081.

Wood, S. N., N. Pya, and B. Säfken (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association 111*(516), 1548–1563.



Location parameters ($n = 1500$)

Dispersion parameters ($n = 1500$)

Figure B.3: Simulation study ($n = 1500$): estimation of the regression parameters in the DALSM model over $S = 500$ replicates: boxplot of the point estimates under a nonparametric (grey) or Normal (white) error term, Relative Efficiency (R.E.) under the working Normality hypothesis, Effective Coverage (E.C.) of 95% credible intervals.

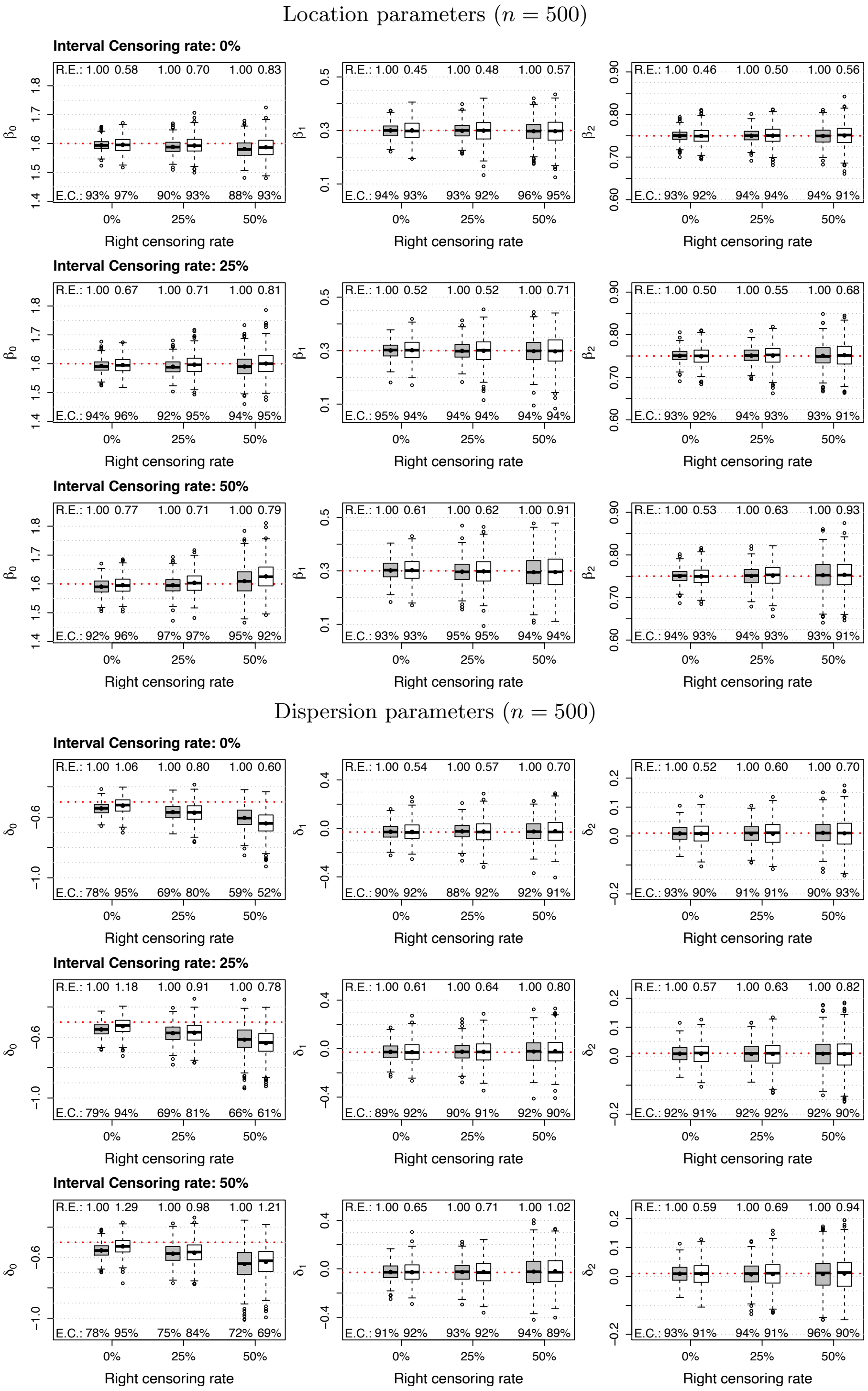


Figure B.4: Simulation study ($n = 500$): estimation of the regression parameters in the DALSM model over $S = 500$ replicates: boxplot of the point estimates under a nonparametric (grey) or Normal (white) error term, Relative Efficiency (R.E.) under the working Normality hypothesis, Effective Coverage (E.C.) of 95% credible intervals.

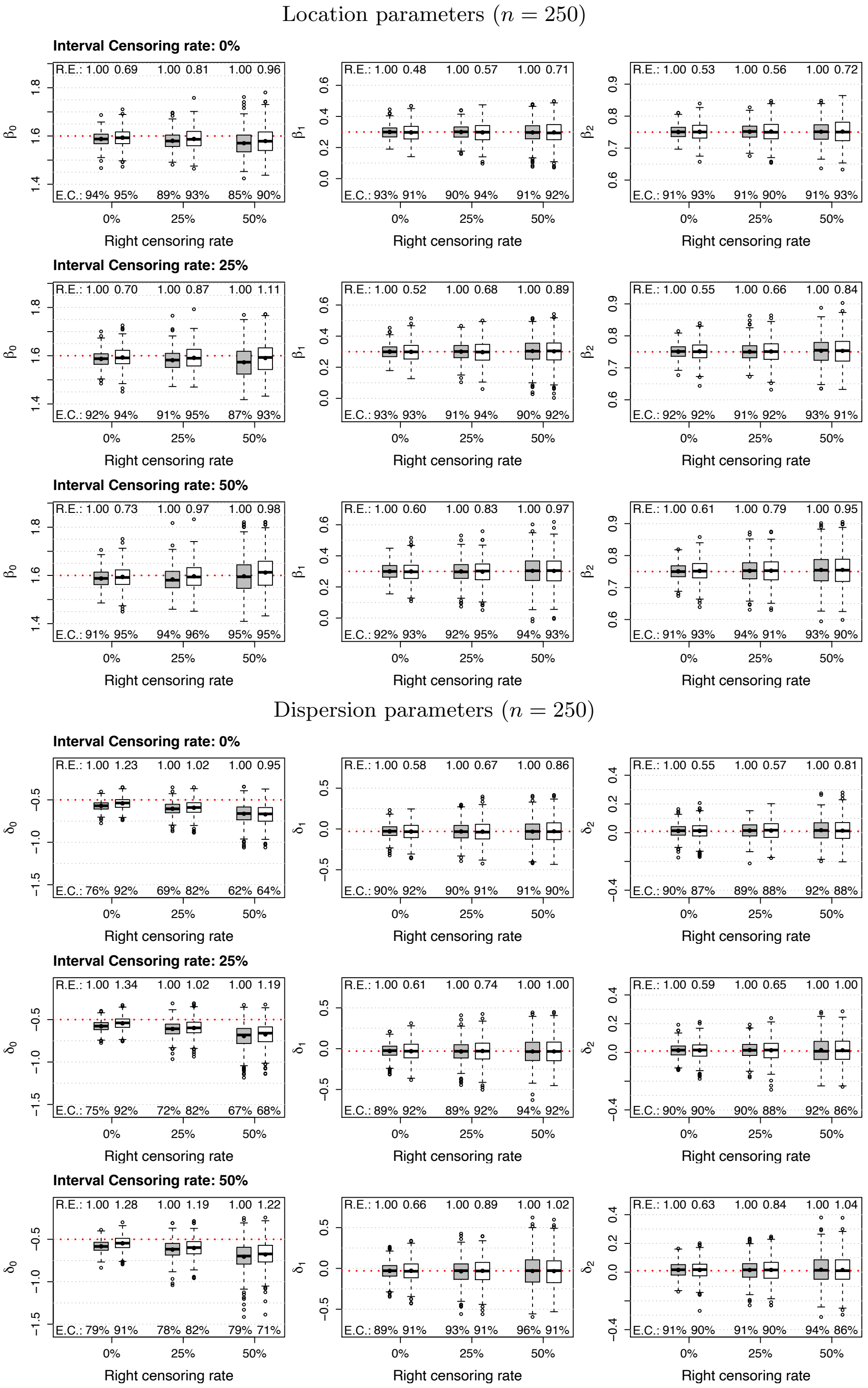


Figure B.5: Simulation study ($n = 250$): estimation of the regression parameters in the DALSM model over $S = 500$ replicates: boxplot of the point estimates under a nonparametric (grey) or Normal (white) error term, Relative Efficiency (R.E.) under the working Normality hypothesis, Effective Coverage (E.C.) of 95% credible intervals.

Table B.2: Simulation study ($n = 1500$): estimation of the additive terms in the DALSM model over $S = 500$ replicates for varying right-censoring (RC) and interval-censoring (IC) rates: Mean absolute bias, Root mean integrated squared error (RMISE), Relative Efficiency with an assumed Normal ($\mathcal{N}$) or nonparametric (NP) error term, Mean effective coverage of 95% credible intervals.

n = 1500			$f_1^\mu(x)$			$f_2^\mu(x)$			$f_1^\sigma(x)$			$f_2^\sigma(x)$		
IC		RC:	0%	25%	50%	0%	25%	50%	0%	25%	50%	0%	25%	50%
0%	MA-Bias	NP	0.003	0.003	0.003	0.002	0.003	0.003	0.002	0.002	0.002	0.011	0.014	0.018
		$\mathcal{N}$	0.003	0.003	0.003	0.003	0.004	0.004	0.002	0.003	0.007	0.014	0.018	0.025
	RMISE	NP	0.016	0.018	0.021	0.015	0.017	0.020	0.026	0.030	0.035	0.037	0.043	0.052
		$\mathcal{N}$	0.022	0.024	0.027	0.021	0.023	0.026	0.035	0.039	0.044	0.049	0.059	0.074
	Rel.Eff.	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\mathcal{N}$	0.508	0.537	0.569	0.496	0.512	0.563	0.522	0.586	0.645	0.557	0.587	0.670
	Coverage	NP	0.959	0.961	0.965	0.961	0.963	0.963	0.960	0.956	0.958	0.942	0.935	0.927
	95% CI	$\mathcal{N}$	0.959	0.961	0.963	0.960	0.958	0.954	0.958	0.960	0.957	0.944	0.934	0.915
25%	MA-Bias	NP	0.003	0.003	0.003	0.003	0.003	0.004	0.002	0.002	0.002	0.013	0.016	0.021
		$\mathcal{N}$	0.003	0.003	0.003	0.004	0.004	0.005	0.002	0.003	0.008	0.015	0.019	0.028
	RMISE	NP	0.017	0.020	0.024	0.016	0.019	0.023	0.027	0.033	0.039	0.040	0.047	0.059
		$\mathcal{N}$	0.023	0.026	0.031	0.022	0.025	0.029	0.037	0.041	0.046	0.051	0.063	0.081
	Rel.Eff.	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\mathcal{N}$	0.543	0.582	0.638	0.535	0.559	0.657	0.549	0.644	0.736	0.593	0.622	0.742
	Coverage	NP	0.961	0.962	0.964	0.957	0.962	0.962	0.961	0.955	0.963	0.937	0.933	0.933
	95% CI	$\mathcal{N}$	0.962	0.962	0.958	0.961	0.957	0.960	0.957	0.960	0.962	0.944	0.933	0.912
50%	MA-Bias	NP	0.003	0.003	0.004	0.003	0.004	0.005	0.002	0.003	0.003	0.015	0.019	0.031
		$\mathcal{N}$	0.003	0.003	0.004	0.004	0.005	0.005	0.002	0.004	0.008	0.017	0.022	0.038
	RMISE	NP	0.019	0.023	0.032	0.018	0.022	0.030	0.029	0.035	0.044	0.044	0.054	0.079
		$\mathcal{N}$	0.024	0.029	0.035	0.024	0.027	0.033	0.038	0.042	0.049	0.055	0.069	0.099
	Rel.Eff.	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\mathcal{N}$	0.585	0.647	0.816	0.582	0.654	0.813	0.590	0.697	0.841	0.613	0.685	0.906
	Coverage	NP	0.960	0.961	0.967	0.958	0.958	0.968	0.962	0.964	0.974	0.939	0.928	0.941
	95% CI	$\mathcal{N}$	0.963	0.958	0.962	0.961	0.959	0.961	0.961	0.963	0.963	0.940	0.924	0.894

Table B.3: Simulation study ($n = 500$): estimation of the additive terms in the DALSM model over $S = 500$ replicates for varying right-censoring (RC) and interval-censoring (IC) rates: Mean absolute bias, Root mean integrated squared error (RMISE), Relative Efficiency with an assumed Normal ($\mathcal{N}$) or nonparametric (NP) error term, Mean effective coverage of 95% credible intervals.

n = 500			$f_1^\mu(x)$			$f_2^\mu(x)$			$f_1^\sigma(x)$			$f_2^\sigma(x)$		
IC		RC:	0%	25%	50%	0%	25%	50%	0%	25%	50%	0%	25%	50%
0%	MA-Bias	NP	0.003	0.003	0.005	0.004	0.006	0.007	0.002	0.003	0.006	0.024	0.033	0.048
		$\mathcal{N}$	0.004	0.005	0.007	0.006	0.007	0.008	0.002	0.004	0.006	0.031	0.040	0.055
	RMISE	NP	0.025	0.029	0.036	0.025	0.028	0.034	0.044	0.051	0.062	0.063	0.076	0.097
		$\mathcal{N}$	0.035	0.039	0.045	0.033	0.036	0.042	0.059	0.064	0.073	0.084	0.098	0.120
	Rel.Eff.	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\mathcal{N}$	0.506	0.554	0.624	0.554	0.576	0.619	0.577	0.627	0.719	0.538	0.619	0.702
	Coverage	NP	0.958	0.960	0.959	0.952	0.955	0.954	0.947	0.946	0.945	0.919	0.895	0.863
	95% CI	$\mathcal{N}$	0.953	0.955	0.951	0.955	0.958	0.952	0.953	0.955	0.950	0.904	0.894	0.868
25%	MA-Bias	NP	0.003	0.004	0.006	0.005	0.007	0.008	0.002	0.004	0.007	0.028	0.039	0.063
		$\mathcal{N}$	0.004	0.005	0.007	0.007	0.008	0.009	0.003	0.006	0.007	0.034	0.046	0.067
	RMISE	NP	0.027	0.032	0.043	0.027	0.031	0.040	0.048	0.055	0.070	0.069	0.085	0.117
		$\mathcal{N}$	0.037	0.042	0.050	0.035	0.040	0.047	0.062	0.066	0.079	0.089	0.106	0.134
	Rel.Eff.	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\mathcal{N}$	0.532	0.603	0.729	0.601	0.620	0.739	0.601	0.686	0.786	0.585	0.649	0.826
	Coverage	NP	0.960	0.959	0.963	0.948	0.954	0.958	0.948	0.952	0.959	0.905	0.890	0.857
	95% CI	$\mathcal{N}$	0.955	0.958	0.954	0.957	0.955	0.954	0.952	0.954	0.951	0.897	0.878	0.854
50%	MA-Bias	NP	0.003	0.005	0.009	0.006	0.008	0.010	0.002	0.005	0.010	0.033	0.048	0.095
		$\mathcal{N}$	0.004	0.007	0.009	0.007	0.009	0.010	0.002	0.006	0.010	0.038	0.054	0.088
	RMISE	NP	0.031	0.037	0.054	0.030	0.036	0.052	0.051	0.062	0.083	0.078	0.098	0.159
		$\mathcal{N}$	0.040	0.046	0.057	0.037	0.044	0.054	0.064	0.071	0.084	0.095	0.118	0.160
	Rel.Eff.	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\mathcal{N}$	0.578	0.676	0.913	0.634	0.689	0.921	0.640	0.771	0.960	0.670	0.711	0.969
	Coverage	NP	0.960	0.964	0.966	0.957	0.960	0.963	0.951	0.957	0.978	0.891	0.888	0.856
	95% CI	$\mathcal{N}$	0.951	0.960	0.954	0.961	0.956	0.951	0.953	0.952	0.954	0.892	0.871	0.818

Table B.4: Simulation study ($n = 250$): estimation of the additive terms in the DALSM model over $S = 500$ replicates for varying right-censoring (RC) and interval-censoring (IC) rates: Mean absolute bias, Root mean integrated squared error (RMISE), Relative Efficiency with an assumed Normal ($\mathcal{N}$) or nonparametric (NP) error term, Mean effective coverage of 95% credible intervals.

n = 250			$f_1^\mu(x)$			$f_2^\mu(x)$			$f_1^\sigma(x)$			$f_2^\sigma(x)$		
IC		RC:	0%	25%	50%	0%	25%	50%	0%	25%	50%	0%	25%	50%
0%	MA-Bias	NP	0.004	0.005	0.007	0.006	0.006	0.008	0.004	0.004	0.007	0.046	0.067	0.103
		$\mathcal{N}$	0.006	0.008	0.009	0.008	0.008	0.009	0.003	0.004	0.007	0.059	0.076	0.100
	RMISE	NP	0.035	0.041	0.056	0.035	0.041	0.053	0.062	0.074	0.099	0.098	0.121	0.164
		$\mathcal{N}$	0.047	0.053	0.063	0.047	0.051	0.059	0.082	0.093	0.113	0.124	0.146	0.173
	Rel.Eff.	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\mathcal{N}$	0.569	0.628	0.793	0.548	0.630	0.809	0.582	0.630	0.768	0.608	0.687	0.875
	Coverage	NP	0.951	0.951	0.934	0.946	0.943	0.934	0.943	0.948	0.950	0.851	0.816	0.770
	95% CI	$\mathcal{N}$	0.958	0.951	0.943	0.950	0.947	0.948	0.954	0.944	0.932	0.849	0.826	0.805
25%	MA-Bias	NP	0.005	0.006	0.010	0.007	0.008	0.010	0.006	0.004	0.005	0.053	0.081	0.123
		$\mathcal{N}$	0.007	0.009	0.011	0.008	0.009	0.010	0.005	0.004	0.007	0.064	0.086	0.114
	RMISE	NP	0.039	0.047	0.067	0.039	0.047	0.063	0.067	0.081	0.117	0.107	0.138	0.196
		$\mathcal{N}$	0.049	0.057	0.070	0.050	0.056	0.066	0.086	0.098	0.121	0.132	0.156	0.195
	Rel.Eff.	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\mathcal{N}$	0.643	0.679	0.913	0.586	0.721	0.902	0.611	0.688	0.940	0.640	0.746	0.941
	Coverage	NP	0.949	0.947	0.941	0.949	0.944	0.939	0.945	0.956	0.959	0.838	0.798	0.783
	95% CI	$\mathcal{N}$	0.956	0.943	0.940	0.948	0.946	0.944	0.952	0.945	0.937	0.838	0.818	0.792
50%	MA-Bias	NP	0.005	0.008	0.012	0.008	0.008	0.010	0.007	0.002	0.007	0.062	0.104	0.137
		$\mathcal{N}$	0.008	0.010	0.013	0.009	0.008	0.010	0.005	0.005	0.009	0.071	0.096	0.130
	RMISE	NP	0.044	0.057	0.077	0.042	0.055	0.073	0.072	0.095	0.130	0.120	0.170	0.218
		$\mathcal{N}$	0.054	0.062	0.079	0.054	0.060	0.073	0.090	0.105	0.131	0.143	0.173	0.214
	Rel.Eff.	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\mathcal{N}$	0.666	0.838	0.956	0.611	0.836	0.992	0.640	0.819	0.990	0.673	0.873	1.017
	Coverage	NP	0.956	0.952	0.949	0.950	0.952	0.959	0.953	0.961	0.978	0.826	0.794	0.846
	95% CI	$\mathcal{N}$	0.953	0.949	0.937	0.947	0.948	0.952	0.947	0.939	0.939	0.824	0.807	0.786

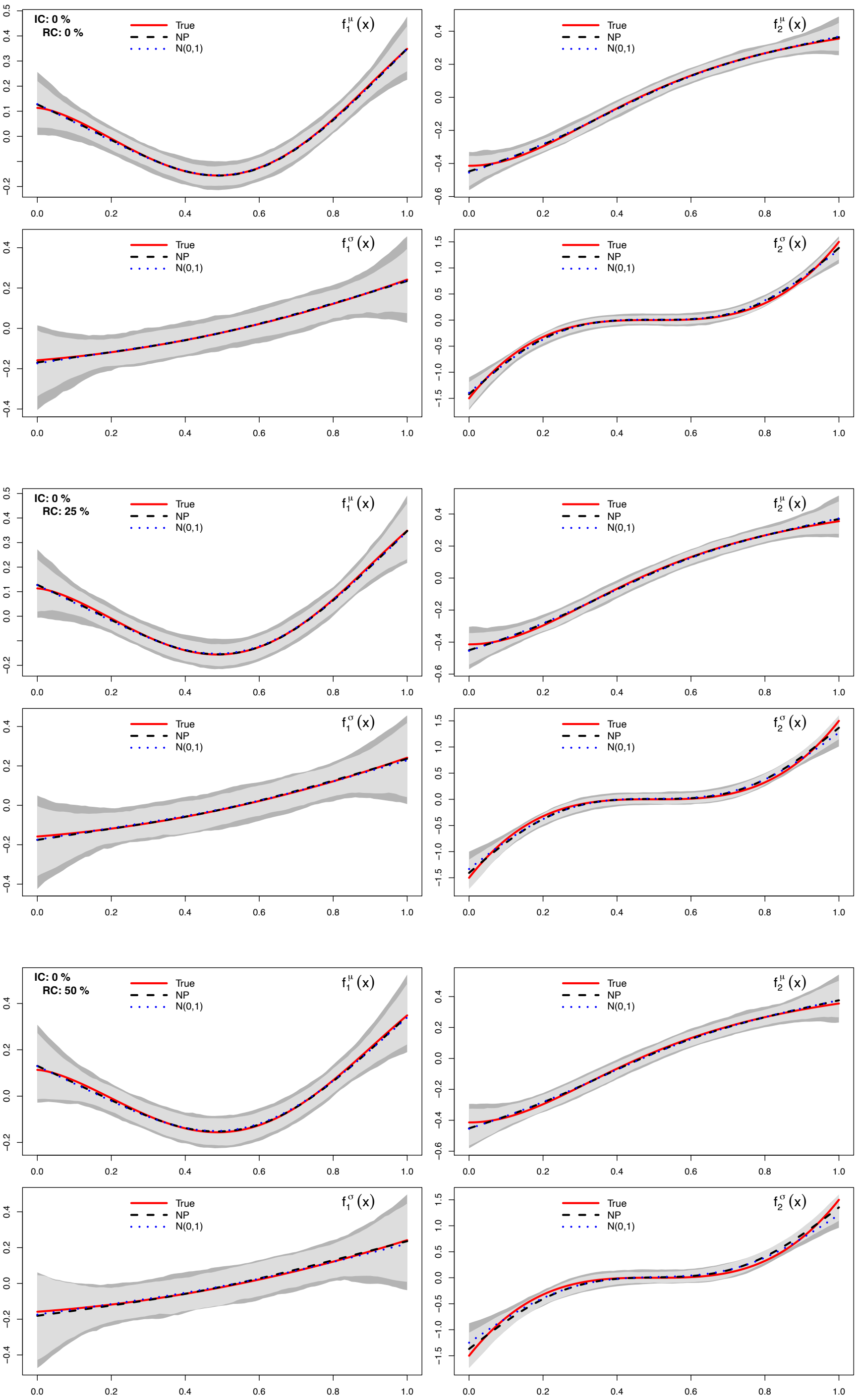


Figure B.6: Simulation study ($n = 500$): averaged estimated additive terms (over $S = 500$ replicates) in the *absence of interval-censoring*, but for increasing right-censoring rates and by assuming a NP (dashed line) or a Normal (dotted line) error term. Envelopes (light grey: NP ; dark grey: Normal) result from consecutive intervals containing 95% of the S estimates for $f_j^{\mu}(x)$ or $f_j^{\sigma}(x)$ with x in $(0, 1)$.

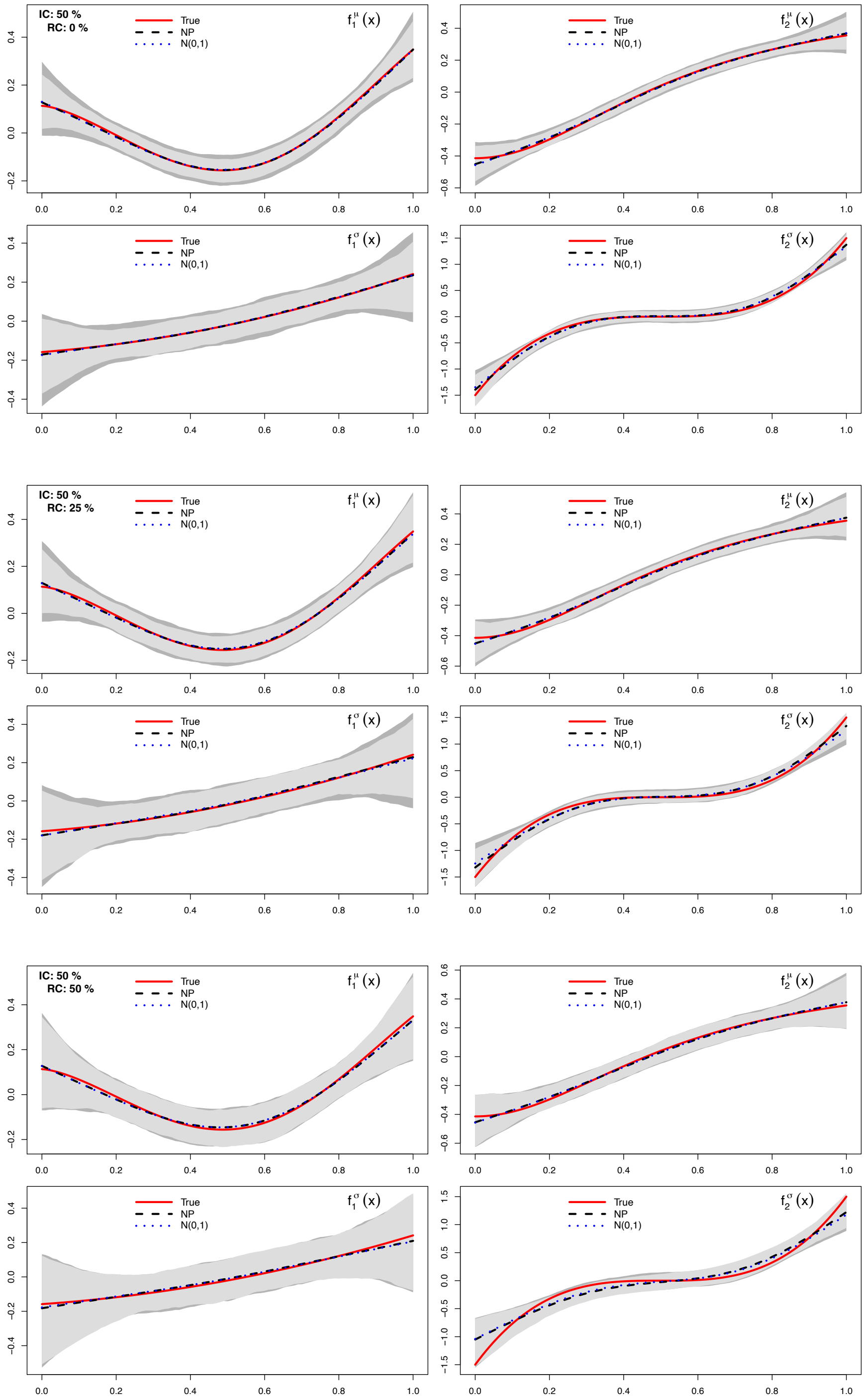


Figure B.7: Simulation study ($n = 500$): averaged estimated additive terms (over $S = 500$ replicates) under a *50% interval-censoring rate* combined with increasing right-censoring rates and by assuming a NP (dashed line) or a Normal (dotted line) error term. Envelopes (light grey: NP ; dark grey: Normal) result from consecutive intervals containing 95% of the S estimates for $f_j^{\mu}(x)$ or $f_j^{\sigma}(x)$ with x in $(0,1)$.

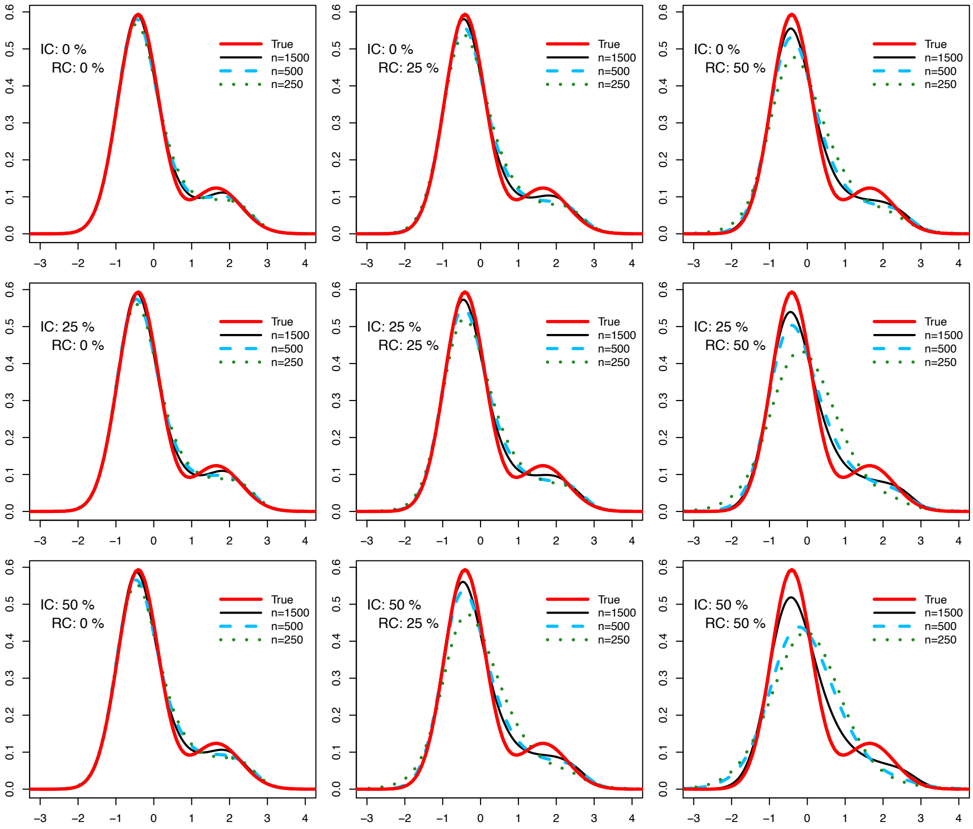


Figure B.8: Simulation study: estimated error densities in the NP-DALSM model (averaged over the $S = 500$ replicates) using a NP error term for different combinations of sample sizes, right- (RC) and interval-censoring (IC) rates.