INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

1051

ADDITIVE LOCATION-SCALE MODELS FOR INTERVAL CENSORED DATA

LAMBERT, P.

This file can be downloaded from
http://www.stat.ucl.ac.be/ISpub

Additive location-scale models for interval censored data

Philippe Lambert *†

December 21, 2010

Abstract

An additive model for the location and dispersion of a continuous response with an arbitrary smooth conditional distribution is proposed. B-splines are used to specify the three components of the model. It can be extended to deal with interval censored data and multiple covariates. As an illustration, the relation between age, the number of years of full-time education and the net income (provided as intervals) available per person in Belgian households is studied from survey data.

Key words: Interval censored data ; additive model ; Smooth distribution.

1 Introduction

Generalized linear models (GLMs) (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) had a major impact on statistical modelling with the synthesis in a single framework of several extensions of normal linear regression models to discrete and positive continuous random variables. After choosing a distribution in the exponential family and a link function, a transform of the mean response could be described as a linear function of covariates with maximum likelihood estimates of the regression coefficients quickly obtained using the iterative weighted least squares algorithm and its implementation in the software GLIM. Extensions of GLMs enable to deal with multi-categorical, multivariate and longitudinal data, random effects (see e.g. Fahrmeir and Tutz, 2001, for details and references), and even allow to jointly model location and dispersion (Jørgensen, 1997). Another relevant extension of GLMs releases the linearity assumption and allows a nonparametric or a flexible parametric specification of the relation between the mean response and continuous covariates: the publication of the book by Hastie and Tibshirani (1990) generated a lot of research about generalized additive models.

*Institut des sciences humaines et sociales, Méthodes quantitatives en sciences sociales, Université de Liège, Liège, Belgium. Email: p.lambert@ulg.ac.be.

†Institut de statistique, biostatistique et sciences actuarielles (ISBA), Université catholique de Louvain, Louvain-la-Neuve, Belgium.

Here, we shall focus on the location-scale model and extend it using some of the preceding principles. If Y is a continuous response and X a set of covariates in $\mathbb{R}^p$, the location-scale model assumes that

$$Y = \mu(X) + \sigma(X)\varepsilon \tag{1.1}$$

where ε is independent of X, $\mu(X)$ denotes the unknown regression surface and $\sigma(X)$ enables to depart from the homoskedastic case. Most often, one further imposes that $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = 1$ such that $\mu(X)$ and $\sigma^2(X)$ are respectively the conditional mean and variance of Y given X.

Lambert and Lindsey (1999) jointly model the location, dispersion, skewness and kurtosis of a continuous response by relating the four parameters of a stable distribution to covariates. That idea was further investigated by Rigby and Stasinopoulos (2001) by replacing the stable distribution with other flexible parametric conditional distributions and regression models by additive ones. Accelerated failure time (AFT) models (see e.g. Collett, 1994) in survival analysis also rely on Eq. (1.1) with log-transformed data. Explanatory variables are assumed to act multiplicatively on the time-scale and for example to affect the progression of a disease. An AFT model can be extended to include random effects (Lambert *et al.*, 2004) similarly to the inclusion of frailty (Vaupel *et al.*, 1979) in the Cox proportional hazards model or to have a flexible conditional distribution (Komárek *et al.*, 2005). Joint models for location and scale were also studied in the nonparametric literature. Hsieh (1996) proposes an empirical process approach to compare two samples of right-censored survival times modelled using an accelerated failure time model with an arbitrary distribution form. Starting from the Beran's estimate $\tilde{F}(y|x)$ of the conditional distribution of Y given X, Van Keilegom and Akritas (1999) estimate $\mu(X)$ and $\sigma(X)$ consistently from right-censored data by discarding the right tail of $\tilde{F}(y|x)$.

The goal of our research is to estimate the location and dispersion of the conditional distribution of Y in Eq. (1.1) using additive models when the distribution of ε has an unknown smooth density and when Y is possibly interval-censored. The plan of the paper is as follows. After a short introduction, Section 2 reminds how a smooth univariate density can be modelled using penalized B-splines (Eilers and Marx, 1996; Lambert and Eilers, 2006, 2009). Section 3 combines the flexible specification for the conditional density with joint additive models for location and dispersion. The roughness penalties and the identification constraint for the pivotal distribution are described in Section 4. Strategies for a joint estimation of the parameters involved in the three levels of the model are proposed in Section 5 together with an algorithm and practical recommendations based on our experience. An extension of the model to interval censored data is presented in Section 6. After a fit of our location-scale additive model to simulated data and an application to the analysis of interval censored income in Sections 7 and 8, we conclude the paper by a discussion in Section 9.

2 Smooth specification of the conditional distribution

Assume that one wants to specify how the conditional distribution of a single continuous random variable Y changes with a set of continuous covariates $X = (X_1, \ldots, X_p)' \in \mathbb{R}^p$. Denote by $f_{Y|X}(y|x)$ the corresponding conditional density.

Further assume that conditionally on $X = x$, the distribution of $\varepsilon = (Y - \mu_x)/\sigma_x$ does not depend on x for suitably specified conditional location and dispersion parameters, μ_x and σ_x. One can find an interval $\mathcal{S} = (s_1, s_2)$ such that the support of ε is most likely included in it. For example, if μ_x and σ_x^2 denote the first two conditional moments, then, by Chebyshev's theorem, one knows that $\Pr(|\varepsilon| \leq k) \geq 1 - k^{-2}$ for $k > 0$. In most practical situations, that probability is far above that lower bound. Therefore, taking $s_2 = -s_1 = 6$ (say) usually provides satisfactory results. Of course, based on contextual information (such as positive skewness of the response) or historical data, more ad-hoc choices could be made.

Consider now a cubic B-splines basis $\{b_k(\cdot) : k = 1, \ldots, K\}$ associated to a large number (20, say) of equidistant knots on $\mathcal{S}$. Let $\{\mathcal{J}_j : j = 1, \ldots, J\}$ be a partition of $\mathcal{S}$ into a large number (100, say) of consecutive bins of equal width Δ with midpoints $u_{j=1}^J$. Let $[B]_{jk} = b_k(u_j)$ be the $J \times K$ matrix of cubic B-splines evaluated the small bin midpoints $u_j (j = 1, \ldots, J)$. Then, the probability to observe ε in $\mathcal{J}_j$ could be specified as

$$\int_{\mathcal{J}_j} f_\varepsilon(e)de = \pi_j = \frac{\exp([B\boldsymbol{\phi}^*]_j)}{\sum_{\ell=1}^L \exp\left([B\boldsymbol{\phi}^*]_\ell\right)} \approx f_\varepsilon(u_j)\Delta \tag{2.1}$$

where $\phi^* = (\phi_1^*, \ldots, \phi_K^*)' \in \mathbb{R}^K$ is a vector of spline parameters.

As $\pi_j(\phi^* + c) = \pi_j(\phi^*)$ for any scalar c, one should constrain ϕ^* for identifiability. We suggest to work with ϕ,

$$\phi_k = \phi_k^* - \log\left(\sum_{j=1}^J \exp\left([B\boldsymbol{\phi}^*]_j\right)\right),$$

such that

$$\pi_j = \exp\left([B\boldsymbol{\phi}]_j\right).$$

For more details on that strategy for estimating a smooth density, we refer to Eilers and Marx (1996), Lambert and Eilers (2006, 2009) and to Section 4 and 5 for aspects specific to the additive location-scale model.

3 Additive models for location and dispersion

Denote by $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ the conditional location and dispersion of Y given $\mathbf{x}$. Depending on the modelling purposes, $\mu(\mathbf{x})$ could be the conditional mean, median or even mode of the response. Likewise, $\sigma(\mathbf{x})$ could be the conditional standard deviation, interquartile range or median absolute deviation, etc.

Assume the following additive model for the conditional location of Y_i given $(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)$:

$$\mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu) = [\boldsymbol{\mu}(\mathbf{x}^\mu, \mathbf{z}^\mu)]_i = \sum_{j=1}^{J_1} f_j^\mu(x_{ij}^\mu) + \left(\beta_0^\mu + \sum_{j=1}^{p_1} \beta_j^\mu z_{ij}^\mu\right), \tag{3.1}$$

where i ($i = 1, \ldots, n$) indexes the units of observation, $\mathbf{z}_i^\mu = (z_{i1}^\mu, \ldots, z_{ip_1}^\mu)'$ is a set of p_1 (mostly) categorical covariates, $\boldsymbol{\beta}^\mu = (\beta_0^\mu, \ldots, \beta_{p_1}^\mu)'$ the vector of regression parameters and $\mathbf{x}_i^\mu = (x_{i1}^\mu, \ldots, x_{iJ_1}^\mu)'$ a set of J_1 continuous covariates taking values in $(0, 1)$. The last constraint is not a restriction as any of the original covariates can be relocated and rescaled to meet that requirement.

Similarly, consider the following additive model for the dispersion of Y_i given $(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma)$:

$$\log \sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma) = \sum_{j=1}^{J_2} f_j^\sigma(x_{ij}^\sigma) + \left(\beta_0^\sigma + \sum_{j=1}^{p_2} \beta_j^\sigma z_{ij}^\sigma\right) \tag{3.2}$$

Provided that these are smooth, the functional forms in $\mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)$ and $\sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma)$ can be approximated using a linear combination of the elements of a (large) B-splines basis $\{s_l(\cdot) : l = 1, \ldots, L\}$ on $(0, 1)$ (see Brezger and Lang, 2006, in a GLM setting):

$$f_j^\mu(x_{ij}^\mu) = \sum_{l=1}^{L} s_l(x_{ij}^\mu)\theta_{lj}^\mu \quad ; \quad f_j^\sigma(x_{ij}^\sigma) = \sum_{l=1}^{L} s_l(x_{ij}^\sigma)\theta_{lj}^\sigma.$$

Denote by $\Theta^\mu = \{\theta_{lj}^\mu\} = [\boldsymbol{\theta}_1^\mu, \ldots, \boldsymbol{\theta}_{J_1}^\mu]$ and $\Theta^\sigma = \{\theta_{lj}^\sigma\} = [\boldsymbol{\theta}_1^\sigma, \ldots, \boldsymbol{\theta}_{J_2}^\sigma]$ the $L \times J_1$ and $L \times J_2$ matrices of spline coefficients for the additive terms in the location and dispersion models, by $\{S_{j_1}^\mu\}_{il} = \{s_l^\mu(x_{ij_1})\}$ and $\{S_{j_2}^\sigma\}_{il} = \{s_l^\sigma(x_{ij_2})\}$ the $n \times J_1$ and $n \times J_2$ B-spline matrices for the j_1th and j_2th additive terms in the location and dispersion models respectively, by Z^μ and Z^σ the $n \times (1 + p_1)$ and $n \times (1 + p_2)$ design matrices for the linear part in (3.1) and (3.2) with regression parameters $\boldsymbol{\beta}^\mu = (\beta_0^\mu, \ldots, \beta_{p_1}^\mu)'$ and $\boldsymbol{\beta}^\sigma = (\beta_0^\sigma, \ldots, \beta_{p_1}^\sigma)'$. Equations (3.1) and (3.2) can be rewritten as

$$\boldsymbol{\mu}(\mathbf{x}^\mu, \mathbf{z}^\mu) = [S_1^\mu \ldots S_{J_1}^\mu] \text{ vec}\,(\Theta^\mu) + Z^\mu \boldsymbol{\beta}^\mu \tag{3.3}$$

$$\log \boldsymbol{\sigma}(\mathbf{x}^\sigma, \mathbf{z}^\sigma) = [S_1^\sigma \ldots S_{J_2}^\sigma] \text{ vec}\,(\Theta^\sigma) + Z^\sigma \boldsymbol{\beta}^\sigma \tag{3.4}$$

4 Penalties and identification constraint

4.1 Roughness penalty

The flexibility provided by the large numbers of B-splines, L for the additive terms and K for the pivotal distribution $f_\varepsilon(\cdot)$, can be counterbalanced by a roughness penalty in a frequentist setting (Eilers and Marx, 1996) or a suitable prior in a Bayesian framework (Lang and Brezger, 2004; Jullion and Lambert, 2007). The chosen order for the penalty will depend on the desired limiting behavior for the functionals for large values of the penalty. We suggest to work with

- a 3rd order penalty for the pivotal distribution, yielding a normal distribution at the limit when the penalty parameter τ^ϕ tends to $+\infty$:

$$\text{pen}^\phi(\tau^\phi, \boldsymbol{\phi}) = -0.5\tau^\phi \sum_k (\phi_k - 3\phi_{k-1} + 3\phi_{k-2} - \phi_{k-3})^2 = -0.5\tau^\phi (D^\phi \boldsymbol{\phi})' D^\phi \boldsymbol{\phi} \ ;$$

- a 2nd order penalty for each additive term in the location model, yielding a linear regression model at the limit for the jth additive term ($j = 1, \ldots, J_1$) when the penalty parameter τ_j^μ tends to $+\infty$:

$$\text{pen}_j^\mu(\tau_j^\mu, \boldsymbol{\theta}_j^\mu) = -0.5\tau_j^\mu \sum_l (\theta_{lj}^\mu - 2\theta_{l-1,j}^\mu + \theta_{l-2,j}^\mu)^2 = -0.5\tau_j^\mu (D^\mu \boldsymbol{\theta}_j^\mu)' D^\mu \boldsymbol{\theta}_j^\mu \ ;$$

- a 1st order penalty for each additive term in the dispersion model, yielding at the limit a conditionally homoskedastic model with respect to the jth continuous covariate ($j = 1, \ldots, J_2$) when the penalty parameter τ_j^ϕ tends to $+\infty$:

$$\text{pen}_j^\sigma(\tau_j^\sigma, \boldsymbol{\theta}_j^\sigma) = -0.5\tau_j^\sigma \sum_l (\theta_{lj}^\sigma - 2\theta_{l-1,j}^\sigma + \theta_{l-2,j}^\sigma)^2 = -0.5\tau_j^\sigma (D^\sigma \boldsymbol{\theta}_j^\sigma)' D^\sigma \boldsymbol{\theta}_j^\sigma.$$

In Bayesian terms, it can be translated into prior distributions on the spline coefficients,

$$p(\boldsymbol{\phi}|\tau^\phi) \propto (\tau^\phi)^{K/2} \exp\left(-0.5\tau^\phi \ \boldsymbol{\phi}' P^\phi \boldsymbol{\phi}\right), \tag{4.1}$$

$$p(\boldsymbol{\theta}_j^\mu|\tau_j^\mu) \propto (\tau_j^\mu)^{L/2} \exp\left(-0.5\tau_j^\mu \ (\boldsymbol{\theta}_j^\mu)' P^\mu \boldsymbol{\theta}_j^\mu\right), \ 1 \le j \le J_1 \tag{4.2}$$

$$p(\boldsymbol{\theta}_j^\sigma|\tau_j^\sigma) \propto (\tau_j^\sigma)^{L/2} \exp\left(-0.5\tau_j^\sigma \ (\boldsymbol{\theta}_j^\sigma)' P^\sigma \boldsymbol{\theta}_j^\sigma\right), \ 1 \le j \le J_2 \tag{4.3}$$

where

$$P^\phi = (D^\phi)' D^\phi + \epsilon I_K \ \ ; \ \ P^\mu = (D^\mu)' D^\mu + \epsilon I_L \ \ ; \ \ P^\sigma = (D^\sigma)' D^\sigma + \epsilon I_L,$$

are full-rank matrices thanks to the addition of a small multiple ϵ (10^{-6}, say) of the identity matrix.

4.2 Likelihood and identification penalty

Given $\boldsymbol{\psi} = (\boldsymbol{\beta}^\mu, \Theta^\mu, \boldsymbol{\beta}^\sigma, \Theta^\sigma)$, one can associate to each observation,

$$\{(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu), (\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma), y_i\},$$

a quantity $\varepsilon_i(\boldsymbol{\psi})$ such that

$$\varepsilon_i(\boldsymbol{\psi}) = \frac{Y_i - \mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)}{\sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma)}.$$

If, for given $\boldsymbol{\psi}$, one denotes by $n_j = n_j(\boldsymbol{\psi})$ ($j = 1, \ldots, J$) the number of observed $\varepsilon_i(\boldsymbol{\psi})$'s ($i = 1, \ldots, n$) belonging to bin $\mathcal{J}_j$, then the conditional joint distribution of $(N_1(\boldsymbol{\psi}), \ldots, N_J(\boldsymbol{\psi}))$ is multinomial,

$$(N_1(\boldsymbol{\psi}), \ldots, N_J(\boldsymbol{\psi}))|\boldsymbol{\phi}) \sim \text{Mult}(n; \pi_1(\boldsymbol{\phi}), \ldots, \pi_J(\boldsymbol{\phi})), \tag{4.4}$$

for values of $\pi_j(\boldsymbol{\phi})$ given by Eq. (2.1). Therefore, the log-likelihood will be

$$\log L(\boldsymbol{\psi}, \boldsymbol{\phi}|\mathcal{D}) = \sum_{j=1}^{J} n_j(\boldsymbol{\psi}) \log \pi_j(\boldsymbol{\phi}).$$

where $\mathcal{D}$ stands for the available data. Extra constraints should be expressed on the spline coefficients $\boldsymbol{\phi}$ involved in the pivotal distribution to force the desired interpretation for $\mu_i = \mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)$ and $\sigma_i = \sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma)$. For example, to interpret μ_i and σ_i as the conditional mean and standard deviation of Y_i, one should make sure that the mean and variance of ε_i are 0 and 1 respectively. Using Eq. (2.1), one has

$$\mathbb{E}[\varepsilon_i|\boldsymbol{\phi}] \approx \mu_{\varepsilon_\phi} = \frac{1}{J}\sum_{j=1}^{J} \pi_j(\boldsymbol{\phi}) u_j \quad ; \quad \mathbb{V}[\varepsilon_i|\boldsymbol{\phi}] \approx \sigma^2_{\varepsilon_\phi} = \frac{1}{J}\sum_{j=1}^{J} \pi_j u_j^2 - \mu^2_{\varepsilon_\phi},$$

where u_j is the midpoint of bin $\mathcal{J}_j$. Therefore, one could add a large identifiability penalty to the log likelihood

$$\text{pen}_{id} = -\kappa \left\{ \mu^2_{\varepsilon_\phi} + (\sigma^2_{\varepsilon_\phi} - 1)^2) \right\}, \tag{4.5}$$

to force mean 0 and variance 1 to the fitted pivotal distribution when κ tends to $+\infty$. This suggests working with the penalized log-likelihood

$$\log L_{\text{pen}}(\boldsymbol{\psi}, \boldsymbol{\phi}|\text{data}) = \sum_{j=1}^{J} n_j(\boldsymbol{\psi}) \log \pi_j(\boldsymbol{\phi}) + \text{pen}_{id}. \tag{4.6}$$

Similar expressions can be obtained for other interpretations of μ_i and σ_i, see e.g. Section 8 where they stand for the conditional median and inter-quartile range, respectively.

5 Bayesian inference

5.1 Joint and conditional posteriors

The full Bayesian model is given by (4.4), (4.1), (4.2), (4.3) and large variance prior distributions for $\boldsymbol{\beta}^\mu, \boldsymbol{\beta}^\sigma$ and the penalty coefficients

$$\tau^\phi, \tau^\mu_{j_1}, \tau^\sigma_{j_2} \sim \text{Exp}(b = 10^{-6}) \text{ with } 1 \le j_1 \le J_1, 1 \le j_2 \le J_2, \tag{5.1}$$

where $\text{Exp}(b)$ denotes an exponential distribution with mean b^{-1}. Hence, the log of the joint posterior is simply the sum of (4.6), of the logarithms of (4.1-4.3) and of

$$-b\left(\tau^\phi + \sum_{j_1=1}^{J_1} \tau^\mu_{j_1} + \sum_{j_2=1}^{J_2} \tau^\sigma_{j_2}\right).$$

The conditional posterior distributions for the penalty parameters can be shown to be

$$(\tau^{\phi}|\boldsymbol{\phi},\mathcal{D}) \sim \mathcal{G}\left(1+0.5K,\ b+0.5\ \boldsymbol{\phi}'P^{\phi}\boldsymbol{\phi}\right), \tag{5.2}$$

$$(\tau^{\mu}_{j_1}|\boldsymbol{\theta}^{\mu}_{j_1},\mathcal{D}) \sim \mathcal{G}\left(1+0.5L,\ b+0.5\ (\boldsymbol{\theta}^{\mu}_{j_1})'P^{\mu}\boldsymbol{\theta}^{\mu}_{j_1}\right) \text{ with } 1\leq j_1\leq J_1, \tag{5.3}$$

$$(\tau^{\sigma}_{j_2}|\boldsymbol{\theta}^{\sigma}_{j_2},\mathcal{D}) \sim \mathcal{G}\left(1+0.5L,\ b+0.5\ (\boldsymbol{\theta}^{\sigma}_{j_2})'P^{\sigma}\boldsymbol{\theta}^{\sigma}_{j_2}\right) \text{ with } 1\leq j_2\leq J_2. \tag{5.4}$$

Unfortunately, the conditional posteriors for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are not of a familiar type. Therefore, the Metropolis-within-Gibbs algorithm is used to sample from the joint posterior.

5.2 MCMC algorithm

For the algorithm to work without extra tuning on specific examples, we found that the data needed to be standardized in a first step (see Lang and Brezger, 2004, for a similar recommendation). The response was first relocated and rescaled using the sample mean and standard deviation of the observed y_i's. The continuous covariates were also relocated and rescaled to take values in $(0,1)$.

5.2.1 Reference algorithm

The sampling algorithm is the following. At iteration m, given the state of the chain at the end of the previous iteration:

1. *Univariate Metropolis steps* for $\boldsymbol{\theta}^{\mu}_{j_1}$ in $f^{\mu}_{j_1}(\cdot)$:
   Let ς^{μ}_{G} denote the variance of the partial residuals $\{y_i-(\beta^{\mu}_0+\sum_{j=1}^{p_1}\beta^{\mu}_j z^{\mu}_{ij}) : i=1,\ldots,n\}$ at the start of iteration m. Let $Q^{\mu}_{j_1}=\frac{1}{\varsigma^{\mu}_{G}}(S^{\mu}_{j_1})'S^{\mu}_{j_1}+P^{\mu}+\epsilon I_L$ and let $L^{\mu}_{j_1}(L^{\mu}_{j_1})'$ be the Cholesky decomposition of that matrix. Denote by $\boldsymbol{\vartheta}^{(m)}_0$ the value of $\boldsymbol{\theta}^{\mu}_{j_1}$ at the start of iteration m.
   For ℓ from 1 to L:
   - Generate $\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^{(m)}_{\ell-1}+\sigma^{\mu}_{j_1,\ell}z\ L^{\mu}_{j_1}\boldsymbol{e}_{\ell}$ where $\boldsymbol{e}_{\ell}$ is the ℓth unit vector of length L, $z\sim\mathcal{N}(0,1)$ and $\sigma^{\mu}_{j_1,\ell}$ is tuned during the burnin period (Haario *et al.*, 2001) to ensure the desired acceptance rate (40% say). For identifiability of the additive model, recenter $\boldsymbol{\vartheta}$ to ensure that $\int_0^1 f^{\mu}_{j_1}(x|\boldsymbol{\vartheta})dx=1$.
   - Conditionally on the other parameters, accept (and set $\boldsymbol{\vartheta}^{(m)}_{\ell}=\boldsymbol{\vartheta}$) or reject (and set $\boldsymbol{\vartheta}^{(m)}_{\ell}=\boldsymbol{\vartheta}^{(m)}_{\ell-1}$) that proposal in a Metropolis step.

   The value of $\boldsymbol{\theta}^{\mu}_{j_1}$ at the end of iteration m is $\boldsymbol{\vartheta}^{(m)}_{L}$.

2. *Univariate Metropolis steps* for $\boldsymbol{\beta}^{\mu}$:
   Let ς^{μ}_{F} denote the variance of the partial residuals $\{y_i-\sum_{j=1}^{J_1}f^{\mu}_j(x^{\mu}_{ij}) : i=1,\ldots,n\}$ for values of the spline parameters given by step 1. Let

$Q^\mu = \frac{1}{\varsigma_F^\mu}(Z^\mu)'Z^\mu$ and let $L^\mu(L^\mu)'$ be the Cholesky decomposition of that matrix. Denote by $\boldsymbol{b}_0^{(m)}$ the value of $\boldsymbol{\beta}^\mu$ at the start of iteration m.

For ℓ from 1 to p_1+1:

- Generate $\boldsymbol{b} = \boldsymbol{b}_{\ell-1}^{(m)} + \sigma_{F,\ell}^\mu z\ L^\mu \boldsymbol{e}_\ell$ where $\boldsymbol{e}_\ell$ is the ℓth unit vector of length p_1+1, $z \sim \mathcal{N}(0,1)$ and $\sigma_{F,\ell}^\mu$ is tuned during the burnin period to ensure the desired acceptance rate (40% say).
- Conditionally on the other parameters, accept (and set $\boldsymbol{b}_\ell^{(m)} = \boldsymbol{b}$) or reject (and set $\boldsymbol{b}_\ell^{(m)} = \boldsymbol{b}_{\ell-1}^{(m)}$) that proposal in a Metropolis step.

The value of $\boldsymbol{\beta}^\mu$ at the end of iteration m is $\boldsymbol{b}_{p_1+1}^{(m)}$.

3. *Univariate Metropolis steps* for $\boldsymbol{\theta}_{j_2}^\sigma$ in $f_{j_2}^\mu(\cdot)$:
Let $Q_{j_2}^\sigma = (S_{j_2}^\sigma)'S_{j_2}^\sigma + P^\sigma + \epsilon I_L$ and let $L_{j_2}^\sigma(L_{j_2}^\sigma)'$ be the Cholesky decomposition of that matrix. Proceed as in Step 1 with $Q_{j_2}^\sigma$ and $L_{j_2}^\sigma$ substituted to $Q_{j_1}^\mu$ and $L_{j_1}^\mu$ respectively, $\boldsymbol{\vartheta}_0^{(m)}$ set to the value of $\boldsymbol{\theta}_{j_2}^\sigma$ at the start of iteration m and specific tuning parameters $\sigma_{j_2,\ell}^\mu$ to reach the desired acceptance rate (40% say). The value of $\boldsymbol{\theta}_{j_2}^\sigma$ at the end of iteration m is $\boldsymbol{\vartheta}_L^{(m)}$.

4. *Univariate Metropolis steps* for $\boldsymbol{\beta}^\sigma$:
Let ς_F^σ denote the variance of $\{\log|y_i - \sum_{j=1}^{J_1} f_j^\mu(x_{ij}^\mu)| : i = 1,\ldots,n\}$ for values of the spline parameters given by step 1. Let $Q^\sigma = \frac{1}{\varsigma_F^\sigma}(Z^\sigma)'Z^\sigma$ and let $L^\sigma(L^\sigma)'$ be the Cholesky decomposition of that matrix. Denote by $\boldsymbol{b}_0^{(m)}$ the value of $\boldsymbol{\beta}^\sigma$ at the start of iteration m. Proceed as in Step 2 with p_2, L^σ and ς_F^σ substituted to p_1, L^μ and ς_F^μ respectively. The value of $\boldsymbol{\beta}^\sigma$ at the end of iteration m is $\boldsymbol{b}_{p_2+1}^{(m)}$.

5. *Univariate Metropolis steps* for spline parameters $\boldsymbol{\phi}$ in the density:
Let $Q^\phi = B'B^\sigma + P^\phi + \epsilon I_K$ and let $L^\phi(L^\phi)'$ be the Cholesky decomposition of that matrix. Proceed as in Step 1 with Q^ϕ and L^ϕ substituted to $Q_{j_1}^\mu$ and $L_{j_1}^\mu$ respectively, $\boldsymbol{\vartheta}_0^{(m)}$ set to the value of $\boldsymbol{\phi}$ at the start of iteration m and specific tuning parameters σ_ℓ^ϕ to reach the desired acceptance rate (40% say). The value of $\boldsymbol{\phi}$ at the end of iteration m is $\boldsymbol{\vartheta}_K^{(m)}$.

6. *Gibbs step for the penalty parameters:*
Generate $\tau^\phi, \tau_{j_1}^\mu\ (1 \le j_1 \le J_1), \tau_{j_2}^\sigma\ (1 \le j_2 \le J_2)$ for iteration m from Eq. (5.2-5.4) with spline parameters set at their values from the previous steps.

From the generated chain, one can build point estimates and credible regions for the spline parameters and any derived quantity.

5.2.2 Starting values

Starting values for the chain can be obtained by taking values of ϕ corresponding to a standard normal distribution. The starting values for the other parameters were set as follows:

– $[\boldsymbol{\beta}^{\mu}]_1$ and $[\boldsymbol{\beta}^{\sigma}]_1$ were set to the associated descriptive statistics for the observed response. For example, if one models the median and the log interquartile range (IQR) of the response, then the initial values are simply taken to be the sample median and log sample IQR of $\{y_i : i = 1, \ldots, n\}$.

– The other regression parameters and the spline parameters involved in the additive parts of the model were all set equal to 0.

– The penalty parameters were set equal to a small value (10^{-1}, say).

5.3 Truncation of the penalty priors

It was found in examples that the chains for the penalty parameters related to some additive components could, in rare occasions, produce long series of very large values for $\tau^{\mu}_{j_1}$ or $\tau^{\sigma}_{j_2}$ corresponding to a large penalization of conditional nonlinear or non constant behaviors for $f^{\mu}_{j_1}$ or $f^{\sigma}_{j_2}$ as encouraged by a large 2nd or a 1st order penalty for location and dispersion, respectively. This is not surprising as e.g. a large value for $\tau^{\mu}_{j_1}$ at iteration m tends to produce at iteration $m+1$ a vector $\boldsymbol{\theta}^{\mu}_{j_1}$ corresponding to a nearly linear $f^{\mu}_{j_1}(\cdot)$ functional which in turn tends to produce a large $\tau^{\mu}_{j_1}$ at iteration $m+1$ through Eq. (5.3). While a very large (10^6, say) and a large (10^3, say) value for $\tau^{\mu}_{j_1}$ at iteration m both generate at iteration $m+1$ spline parameters suggesting linearity of the functionals $f^{\mu}_{j_1}(\cdot)$, the chain for $\tau^{\mu}_{j_1}$ tends to stay located at the right tail of the posterior distribution of the penalty during a large number of iterations in the first case but not in the second. Therefore, the priors for the penalty parameters in Eq. (5.1) were truncated to $(0, \tau_{\max})$ where $\tau_{\max}$ denotes a large value (say 10^3) for the penalty.

5.4 The identification penalty

The identification penalty in Eq. (4.5) ensures that the estimated pivotal distribution has location 0 and dispersion 1 when κ tends to $+\infty$. While that strategy works when it comes to compute maximum (penalized) likelihood estimates, it has a strong negative impact on the mixing of the chains generated by the MCMC algorithm. Therefore, we advocate the following strategy:

1. Set κ equal to a moderate value (100, say).

2. Run the Metropolis-within-Gibbs algorithm described above.

3. Discard the burnin iterations, yielding after convergence final chains of length M. For m from 1 to M:

   For values of the parameters at iteration m:

– Define the density $g(\cdot|\boldsymbol{\phi})$ at the bin midpoints u_j by

$$g(u_j|\boldsymbol{\phi}) = \frac{1}{\Delta}\ \frac{\exp([B\boldsymbol{\phi}]_j)}{\sum_{\ell=1}^{L}\exp\left([B\boldsymbol{\phi}]_\ell\right)}$$

– Compute the location λ and dispersion δ of $g(\cdot|\boldsymbol{\phi})$.

– Set $\beta_0^\mu \longleftarrow \beta_0^\mu + \lambda \quad \text{and} \quad \beta_0^\sigma \longleftarrow \beta_0^\sigma + \log\delta$.

– The pivotal distribution with location 0 and dispersion 1 corresponds to the density f_ε given by

$$g(\cdot|\boldsymbol{\phi}) = \frac{1}{\delta}\ f_\varepsilon\left(\frac{\cdot-\lambda}{\delta}\middle|\ \boldsymbol{\phi}\right).$$

– Recompute $\boldsymbol{\phi}$ such that $f_\varepsilon(u_j|\boldsymbol{\phi})\Delta = \exp([B\boldsymbol{\phi}]_j)$ for $1 \le j \le J$.

After these corrections to the reference algorithm (see Section 5.2.1), our experience on different examples showed that trace plots suggest convergence after just a few thousands iterations.

6 Extension to interval censored data

Assume now that the response data take the form of intervals $\{(y_i^L, y_i^U) : i = 1, \ldots, n\}$. Then, these intervals can be relocated and rescaled as (e_i^L, e_i^U) where

$$e_i^L = e_i^L(\boldsymbol{\psi}) = \frac{y_i^L - \mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)}{\sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma)} \quad ; \quad e_i^U = e_i^U(\boldsymbol{\psi}) = \frac{y_i^U - \mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)}{\sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma)}.$$

Assume that the exact unobserved response Y_i is standardized in the same way to obtain ε_i and that ε_i has a conditional distribution independent of the covariates. Consider now the same strategy and notation as in Section 2 to model the corresponding pivotal density $f_\varepsilon(\cdot)$ on interval $\mathcal{S}$. If $\boldsymbol{\xi} = (\boldsymbol{\phi}, \boldsymbol{\psi})$ and if c_{ij} is the proportion of bin $\mathcal{J}_j$ contained in $(\varepsilon_i^L, \varepsilon_i^U)$, then

$$\begin{aligned}\Pr_{\boldsymbol{\xi}}\left(y_i^L < Y_i < y_i^U | \mathbf{x}_i^\mu, \mathbf{z}_i^\mu, \mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma\right) &= \Pr_{\boldsymbol{\xi}}\left(e_i^L < \varepsilon_i < e_i^U\right) = \int_{e_i^L}^{e_i^U} f_\varepsilon(e|\boldsymbol{\phi})de \\ &\approx \sum_j c_{ij}(\boldsymbol{\psi})\pi_j(\boldsymbol{\phi}),\end{aligned}$$

with $\pi_j(\boldsymbol{\phi})$ given in Eq. (2.1). This is the composite link model (Thompson and Baker, 1981; Eilers, 2007) and its application to the estimation of a density from interval censored data (Lambert and Eilers, 2009; Lambert, 2011). It could also be extended to deal with heavily right censored data along the ideas in Çetinyürek and Lambert (2010).

The penalized log-likelihood that one had in Eq. (4.6) for precisely observed responses becomes

$$\log L_{\text{pen}}(\theta|\text{data}) = \sum_{i=1}^{I}\log\left(\sum_{j=1}^{J} c_{ij}\pi_j\right) + \text{pen}_{id}.$$

The derivation of the joint posterior and the strategy for exploring it are unchanged.

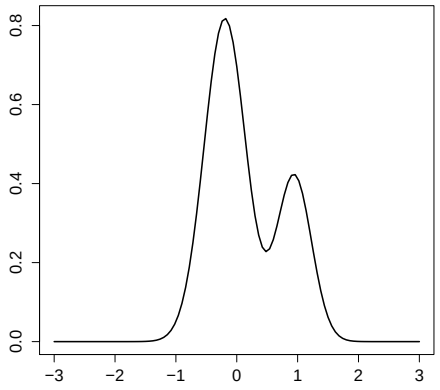

Figure 1: Simulation: conditional density $f_\varepsilon(\varepsilon)$.

7 Simulation study

One hundred datasets $\{(x_i, y_i) : i = 1, \ldots, n\}$ of size $n = 500, 1000, 1500$ were generated from the location-scale model in Eq. (1.1) where X is uniformly distributed on $(0, 1)$, $\mu(x) = \sin(2\pi x), \sigma(x) = 0.79 \exp\{\cos(4\pi x)\}$ and $\varepsilon = 1.758W - 0.657$ with $W = 0.7\ \mathcal{N}\left(-1,\ 0.6^2\right) + 0.3\ \mathcal{N}\left(1,\ 0.5^2\right)$. One can check that ε has median 0 and interquartile range 1, see Fig. 1 for a graphical representation. Therefore, $\mu(x)$ and $\sigma(x)$ are the conditional median and interquartile range of Y given $X = x$, see Fig. 2. One also has $\mathbb{E}(\sigma(X)) = 1$. A scatterplot of one generated dataset of size $n = 500$ is provided on Fig. 3. The points on Fig. 3 are the midpoints of the observed interval data $(y_i - 0.5h\ \sigma_{x_i}, y_i + 0.5h\ \sigma_{x_i})$ of average width h (taken to be 0.1, 0.5 or 1.0 in the design of the simulation study). That makes $100 \times 3 \times 3$ datasets for which $\mu(x)$, $\sigma(x)$ and $f_\varepsilon(\varepsilon)$ were estimated using the Bayesian strategy exposed in Section 5.

The estimated conditional median, interquartile range and pivotal density for each of the 100 simulated datasets are plotted as grey curves on Fig. 4, 5 and 6, respectively ; the dashed curves represent the true functional. One can see that the estimation of the conditional median and IQR is performant whatever the sample size n and the mean width h of the interval data. Not surprisingly, the uncertainty decreases as n increases. The pivotal density and its bimodality is also clearly detected. The quality of the reconstruction deteriorates as h increases and is least satisfactory when $h = 1.0$.

Of course, in practice, if the available information is too sparse to provide fine estimates of some of the components, the penalties will guide us, at the limit, towards a classical linear regression setting with a normal pivotal distribution.

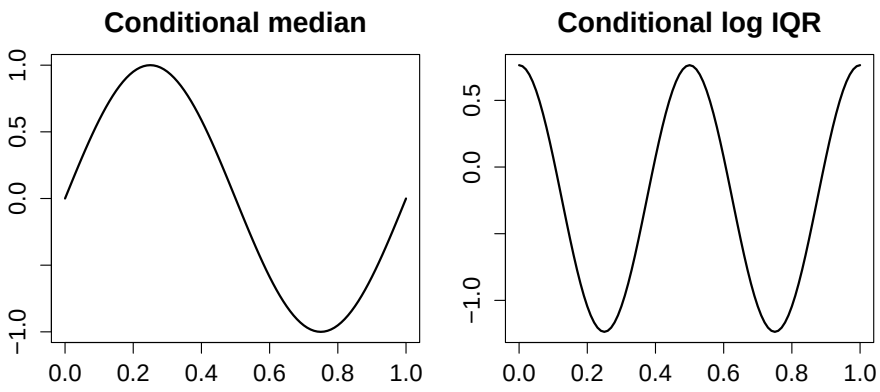

Figure 2: Simulation: conditional median $\mu(x)$ and interquartile range $\sigma(x)$.

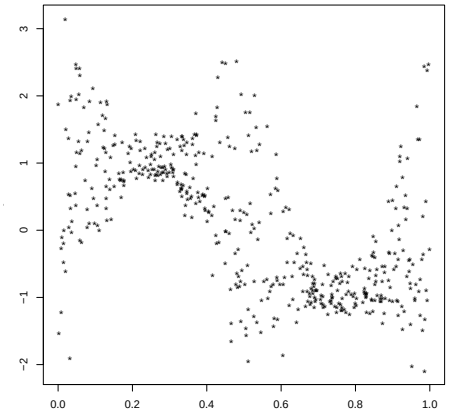

Figure 3: Simulation: scatterplot of $n = 500$ simulated data.

8 Application

The data of interest were collected in 2006 through the European Social Survey, Round 3 (ESS, 2006). This is an academically-driven social survey designed to study the attitudes, beliefs and behaviour patterns of diverse populations in more than 30 nations in Europe. We are specifically interested in the money available per person in Belgian households for respondents aged between 25 and 75 who studied between 8 and 20 years. The ESS provides the net monthly income (in €) of the household ($n = 1103$) reported in one of the following intervals: 1: < 150 2: $[150, 300[$, 3: $[300, 500[$, 4: $[500, 1.000[$, 5: $[1.000, 1.500[$, 6: $[1.500, 2.000[$, 7: $[2.000, 2.500[$, 8: $[2.500, 3.000[$, 9: $[3.000, 5.000[$, 10: $[5.000, 7.500[$, 11: $[7.500, 10.000[$, 12: ≥ 10.000€. The available income per person (also named the *equivalised household income*) is obtained by dividing the household net income by the household weight. That weight is calculated using the OECD-

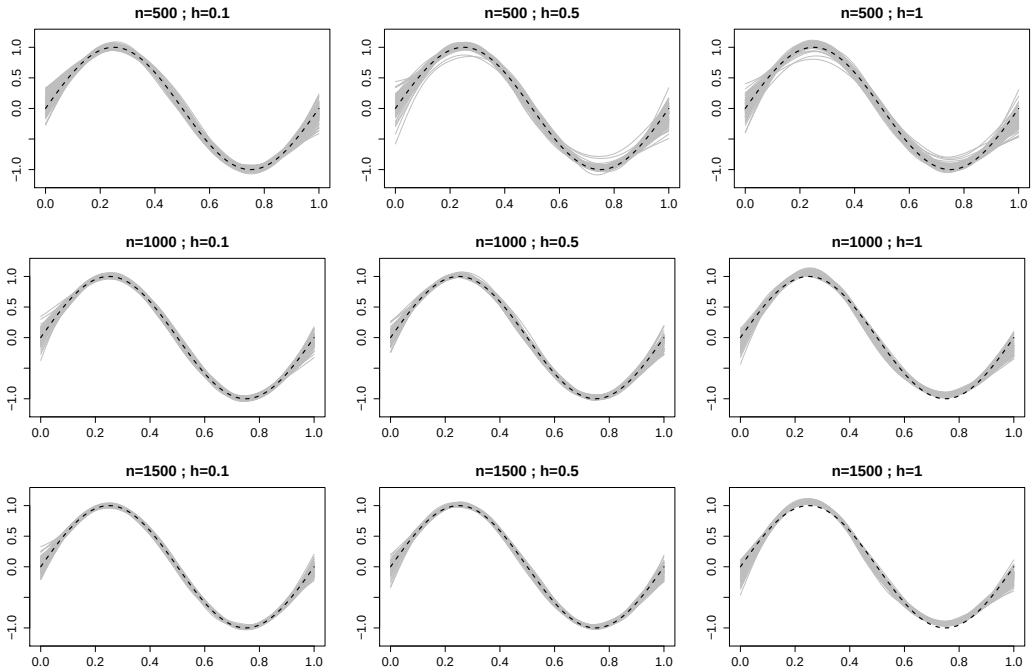

Figure 4: Simulation: estimated conditional median (one grey curve per dataset) for sample sizes 500, 1000 or 1500 with interval data of average width $h = 0.1$, 0.5 or 1.0. The dashed line is the true conditional median.

modified scale first proposed by Hagenaars *et al.* (1994) and advocated since the late 1990s by the Statistical Office of the European Union (EUROSTAT). This scale assigns a value of 1 to the household head, of 0.5 to each additional adult member and of 0.3 to each child (less than 14 years old). Therefore, our data consists of the preceding intervals each time divided by the household weight. We have studied its relation with age (48.5 ± 13.4 years) and the number of years of full-time education completed (12.7 ± 2.8 years) by the respondent.

We assume an additive location-scale model for the available income per person, Y, cf. Eq. (1.1), and force ε to have median 0 and inter-quartile range 1. The continuous covariates are `age` and `educ` such that $\mathbf{x}_i = (\text{age}_i, \text{educ}_i)$. Thus, $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ must be interpreted as the conditional median and inter-quartile range of Y given $X = \mathbf{x}$, respectively.

Using the algorithm described in Section 5.2, a final chain of length $50,000$ was run after a burnin of $5,000$ iterations for paramaters $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$. The `age` and `educ` components in the additive models for the conditional median and (log-) inter-quartile range of the income available per person are reported in Fig. 7. These functional components are added to the reference values β_0^μ and β_0^σ in Eqs. (3.1) and (3.2), see Table 1 for their estimated posterior mean and 95% highest posterior density (HPD) interval.

It suggests a significant increase (at a given age) of the median and of

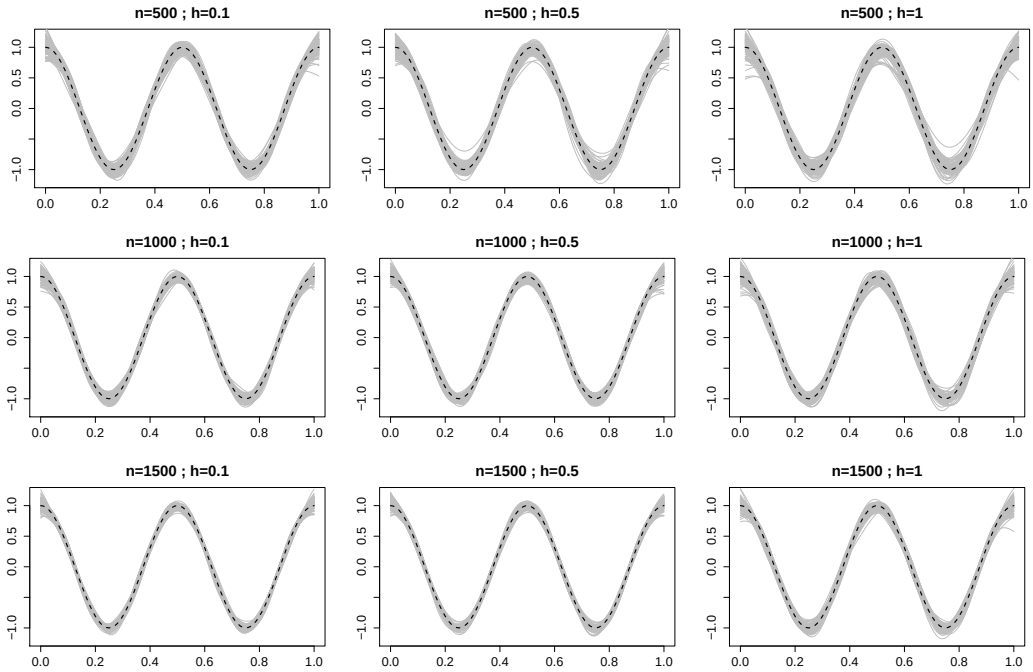

Figure 5: Simulation: estimated conditional log IQR (one grey curve per dataset) for sample sizes 500, 1000 or 1500 with interval data of average width $h = 0.1$, 0.5 or 1.0. The dashed line is the true conditional log IQR.

Parameter	Mean	95% HPD interval
β_0^μ	1461	(1410, 1513)
β_0^σ	6.75	(6.69, 6.80)

Table 1: Income dataset: estimates of the posterior mean and of the 95% HPD interval for the regression intercepts in the additive location-scale model.

log IQR of the net income available per person with the number of years of education. An increase of these quantities with age (for a given education level) is also visible till the late fifties where both the conditional median and log IQR start to decrease. It corresponds to the age where Belgians typically leave the work market (on average at about 61) before the legal age of 65. Indeed, one can retire earlier at 60 after 35 years of contribution to the pension system. The estimated pivotal distribution can be found on Fig. 8: it is right skewed with a long right tail as expected with income data.

From the fitted model, one can compute the conditional probability to live below the poverty treshold. That treshold is commonly defined as 60% of the median equivalised household income. In Belgium, it was 772€ in 2006. If Y

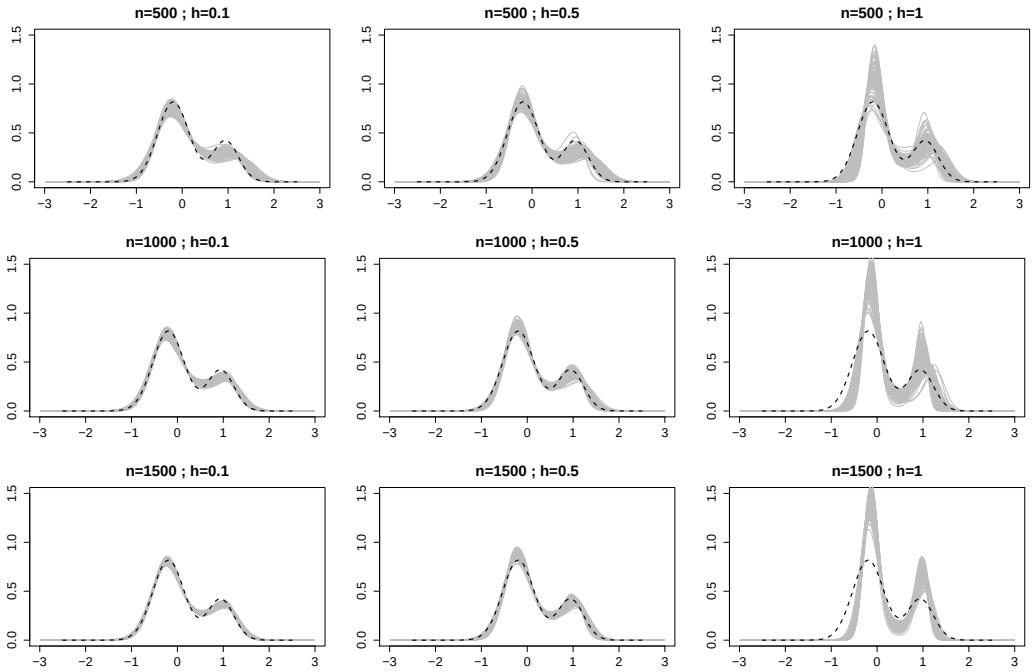

Figure 6: Simulation: estimated pivotal density (one grey curve per dataset) for sample sizes 500, 1000 or 1500 with interval data of average width $h = 0.1$, 0.5 or 1.0. The dashed line is the true pivotal density $f_\varepsilon(\varepsilon)$.

denotes the equivalised household income, then the probability of interest is

$$\Pr(Y \leq 772|X = \mathbf{x}) = \Pr\left(\varepsilon \leq \frac{772 - \mu(\mathbf{x})}{\sigma(\mathbf{x})}\right) = \int_{-\infty}^{\frac{772-\mu(\mathbf{x})}{\sigma(\mathbf{x})}} f_\varepsilon(e)de.$$

It was computed for a grid a values for `age` and `educ`, see Fig. 9. It suggests that education is the best insurance against poverty with (roughly) an estimated risk of more than 10% or 20% for persons who studied for less than 14 or 12 years, respectively. The low-educated young and old persons are particularly exposed with an estimated risk over 30%.

9 Discussion

Additive models were proposed to model the location and the dispersion of a continuous response with an unknown smooth conditional distribution. B-splines were used to specify all the functional components in the model, including the pivotal density. These were estimated jointly in a Bayesian framework where all sources of uncertainty are taken into account. Credible regions for any quantity of interest can be built from the generated MCMC chains.

The proposed tool can deal with interval censored data. It relies upon the results in Lambert and Eilers (2009) and in Lambert (2011) where densities

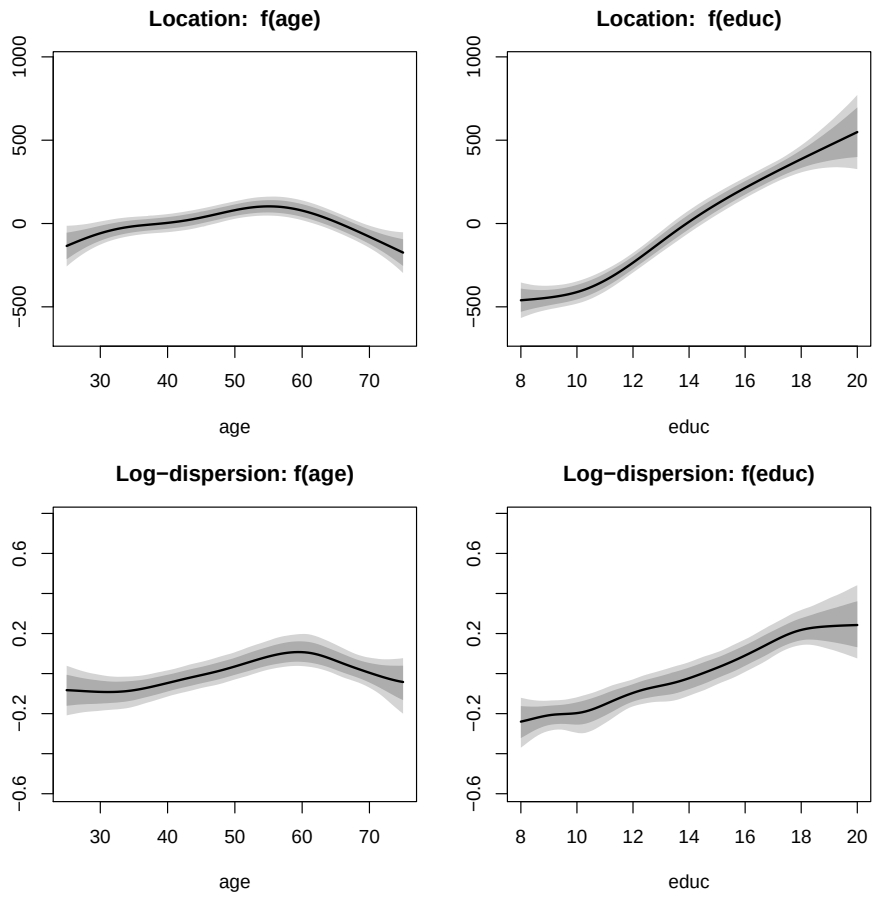

Figure 7: Income data: `age` and `educ` components (in €, solid line) in the additive models for the median and for the log interquartile range (solid line) ; grey regions delimit pointwise 80% (dark grey) and 95% (light grey) credible intervals.

were estimated from univariate or bivariate grouped data. Simulations suggest that a reliable estimation of the additive components can be obtained even with modest sample sizes. A good identification of the pivotal distribution is more dependent on the width of the interval data. The suggested method enables not only to regress moments but also quantiles on covariates and to estimate (in a single step) the response distribution conditionally on these.

Several extensions of that model and further research are desirable. It includes regression on interval censored covariates, model selection, goodness-of-fit tools for interval censored responses, random effects and extension to spatial data. We are currently working on several of these aspects and hope to be able to report soon in separate publications.

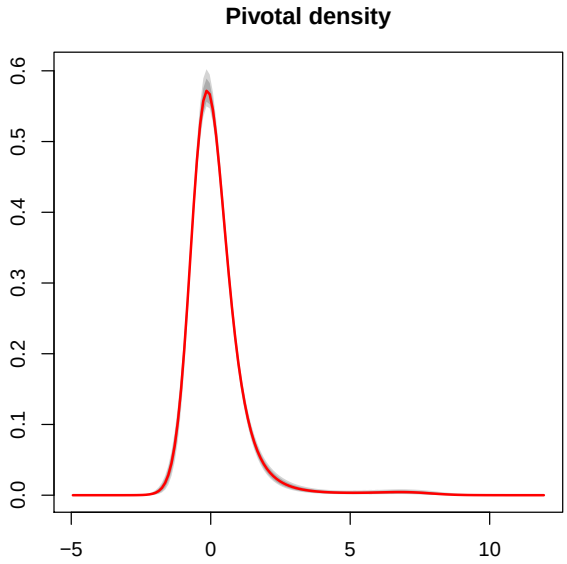

Figure 8: Income data: fitted pivotal density in the location-scale model ; grey regions delimit pointwise 95% credible intervals.

Acknowledgments

Financial supports from the FSR research grant nr. FSRC-08/42 from the University of Liège and of the IAP research network nr. P6/03 of the Belgian government (Belgian Science Policy) are gratefully acknowledged.

References

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

Çetinyürek, A. and Lambert, P. (2010). Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines. *Statistics in Medicine*. DOI: 10.1002/sim.4081.

Collett, D. (1994). *Modelling Survival Data in Medical Research*. Chapman & Hall, London.

Eilers, P. H. C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, **7**, 239–254.

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

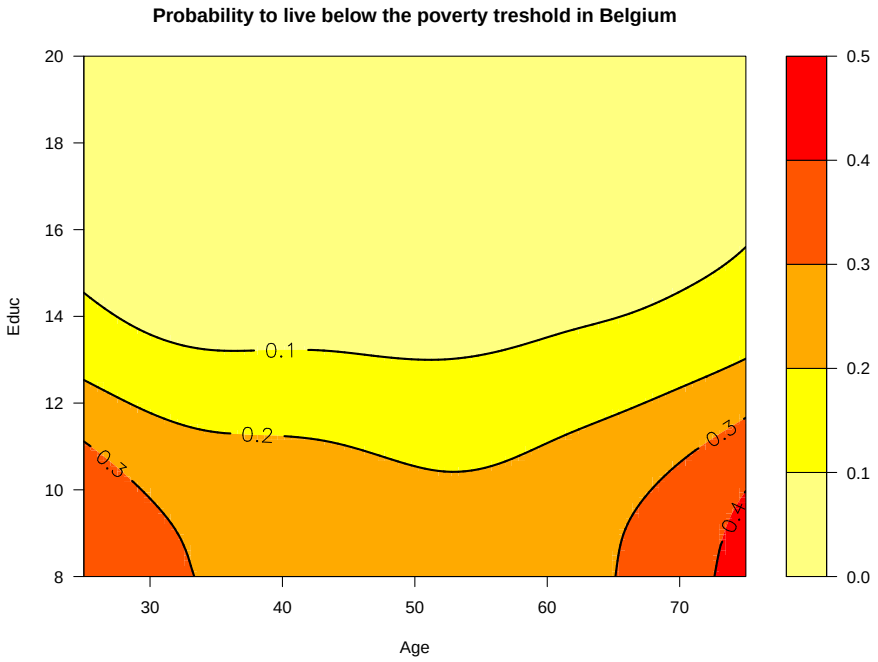

Figure 9: Income data: estimated probability to live below the poverty treshold in 2006 in Belgium.

ESS (2006). European Social Survey Round 3: data file edition 3.2. *Norwegian Social Science Data Services, Norway.* http://www.europeansocialsurvey.org/.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models (2nd edition)*. Springer-Verlag.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, **7**, 223–242.

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

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. Chapman & Hall, London.

Hsieh, F. (1996). Empirical process approach in a two-sample location–scale model with censored data. *The Annals of Statistics*, **24**(6), 2705–2719.

Jørgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall, London.

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive bayesian P-splines models. *Computational Statistics and Data Analysis*, **51**, 2542–2558.

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

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

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

Lambert, P. and Eilers, P. H. C. (2006). Bayesian multidimensional density smoothing. In *Proceedings of the 21st International Workshop on Statistical Modelling*, pages 313–320, Galway.

Lambert, P. and Lindsey, J. (1999). Analysing financial returns using regression based on non-symmetric stable distributions. *Applied Statistics*, **48**, 409–424.

Lambert, P., Collett, D., Kimber, A., and Johnson, R. (2004). Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in Medicine*, **23**(20), 3177–3192.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd Edition*. Chapman & Hall / CRC.

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

Rigby, R. and Stasinopoulos, D. (2001). Generalized additive models for location scale and shape. *Applied statistics*, **54**(3), 1–38.

Thompson, R. and Baker, R. J. (1981). Composite link functions in generalized linear models. *Applied Statistics*, **30**, 125–131.

Van Keilegom, I. and Akritas, M. (1999). Transfer of tail information in censored regression models. *The Annals of Statistics*, **27**(5), 1745–1784.

Vaupel, J., Manton, K., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**(3), 439–454.