Chilean Journal of Statistics
Vol. 8, No. 2, September 2017, 3-23

Analysis of Single-Index Models with Scale Mixture of Normals Errors by Using Bayesian P-Splines

Marcelo M. Taddeo[1,*], and Pedro A. Morettin[2]

[1]Department of Statistics, Federal University of Bahia, Salvador, Brazil,
[2]Department of Statistics, University of São Paulo, São Paulo, Brazil,

(Received: February 1, 2017 · Accepted in final form: August 7, 2017)

Abstract

We consider the estimation of the link function as well as the parameter vector of a single-index model under a purely Bayesian perspective by using P-splines and assuming errors distributed according to a scale mixture of Normals. This approach includes, among others, the Gaussian distribution itself and the Student-t distribution. We have made explicit all the details of the MCMC algorithm used to sample the parameters of interest according to the posterior distribution, including all the posterior full conditionals and Metropolis-Hastings steps. The results of the suggested procedure has been tested and shown through simulation study and an application to real data.

Keywords: Single-Index Model · P-Splines · MCMC · Scale Mixtures of Normals.

Mathematics Subject Classification: 62G08.

1. Introduction

Single-Index models are a way to overcome the well-known Curse of Dimensionality, typical of nonparametric multivariate models. In this paper, the noise was allowed to follow a scale mixture of Gaussians which represents a broader class of distributions than the usual class of Normal distributions. In fact, besides the Gaussian distribution itself, such class includes, among others, the Student's t distribution and the Stable family. All elements of this class, with the exception of the Normal distribution, are heavy tailed and, therefore, they are useful in making models resistant to outliers and extreme values.

The single-index model (SIM) is given by

$$y_t = g(\boldsymbol{\beta}'\boldsymbol{x}_t) + \delta\epsilon_t, \tag{1}$$

for $t = 1, \ldots, T$, where g is known as the link function and the parameter vector $\boldsymbol{\beta}$ is known as the index vector. We want to estimate such components under the assumption that ϵ_t follows a scale mixture of Normals and that $\boldsymbol{x}_t$ is a p-dimensional input vector by using Bayesian P-splines. Since $\boldsymbol{\beta}$ is only identifiable up to a multiplicative constant, we impose that $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$.

Single-index models have been approached in several ways, especially in cases where the noise is Gaussian. The classical approaches often use splines or kernel and consist usually

[*]Corresponding author. Email: marcelo.magalhaes@ufba.br

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

of a two-step procedure, where the vector index is estimated in one step, and the link function is estimated in the next one. See, for example, Yu *et al.* (2002) and the references therein.

Despite the consistency of the estimators, a drawback that has persisted in the classical approach is the numerical instability and, as noted by Antoniadis *et al.* (2004), a Bayesian approach offers the hope of more stable estimates. They use a hybrid approach in which the parameters are generated via MCMC (Bayesian approach) and one of the hyperparameters is obtained via regularization (classic approach). Although much of our attention is in the combination of a Bayesian approach with the application of splines to Single-Index models, other different approaches related to them are available and we cite, as an example, Park *et al.* (2005) who estimate the target function by using wavelets. Even though such works are related to this article, we believe to be the first to enable the simultaneous use of a general class of distributions (the scale mixture of Gaussians) and the function approximation method via P-splines to calibrate a SIM under a purely Bayesian setup.

This paper is organized as follows. In sections 2 and 3, we briefly review and introduce the concepts used here. In section 4, we define the priors of the parameters of interest. In section 5, we derive the posterior full conditional distributions and the Metropolis-Hastings steps used in the MCMC algorithm. Finally, in section 6, we illustrate the proposed method through simulation and an application to real data.

2. Observation Model

A random variable Y is said to follow a scale mixture of Normals (SMN) if it can be written as $Y = Z/\sqrt{\sigma}$, with $Z \sim \mathcal{N}(0,1)$ and σ being any positive (continuous or discrete) random variable. The distribution H associated to σ is called mixture distribution and it determines the particular distribution X will follow. Whenever H is absolutelly continuous, the probability density functions of X and σ (denoted by h) will be connected by the following expression

$$p(y) = \int_0^\infty \sigma^{1/2}\phi(\sigma^{1/2}y)h_{\boldsymbol{\zeta}}(\sigma)d\sigma,$$

where ϕ is the pdf of the standard normal distribution and $\boldsymbol{\zeta}$ stands for the vector of parameters and possibly hyperparameters associated to the mixture distribution. It is not the scope of this paper to give a detailed account of SMN distributions, so for a more comprehensive treatment on this subject we refer to Andrews *et al.* (1974) and Fernandez *et al.* (2000). It should be enough to say that such class is quite large in the sense that it contains all continuous and symmetric distributions, see Fang *et al.* (1989).

From a distributional perspective, we can write

$$\begin{aligned} y_t|\boldsymbol{x}_t;\sigma_t &\sim \mathcal{N}\left(g(\boldsymbol{\beta}'\boldsymbol{x}_t), \frac{\delta^2}{\sigma_t}\right), \\ \sigma_t &\sim h_{\boldsymbol{\zeta}}, \end{aligned} \tag{2}$$

where y_t is the dependent variable in model (1) and $\boldsymbol{\zeta}$ is just a parameter vector which determines the prior distribution of $\boldsymbol{\sigma}$. Here, σ_t is not observable and so it plays the role of a latent variable. From (2), it follows that

$$p(y_t|\boldsymbol{x}_t) = \int_0^\infty \frac{\sigma_t^{1/2}}{\delta}\phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t - g(\boldsymbol{\beta}'\boldsymbol{x}_t))\right)h_{\boldsymbol{\zeta}}(\sigma_t)d\sigma_t,$$

which is just a scaled and shifted version of $p(x)$ as given above.

3. Splines Setup

In order to estimate the link function, we follow Wang *et al.* (2009) and do not approximate g directly, but instead a variant $G \equiv g \circ F_p^{-1}$, where F_p is the rescaled centered Beta$((p+1)/2, (p+1)/2)$ c.d.f[1],

$$F_p(\nu) = \int_{-1}^{\nu/a} \frac{\Gamma(p+1)}{\Gamma\left(\frac{p+1}{2}\right)^2 2^p}(1-v^2)^{\frac{p-1}{2}}dv, \quad \nu \in [-a, a],$$

and we refer to G as the link function. Before justifying the choice for this transformation, assume the pdf p_X of the input random vector $\boldsymbol{X}_t$ satisfies the following assumptions.

Assumption 1: The support of the pdf p_X, denoted by $\mathcal{X}$, is a compact subset of R^p.

Assumption 2: There exist constants $0 < \kappa_\ell \leq \kappa_u < \infty$ and a closed ball in $B = \{\boldsymbol{x} \in \mathrm{R}^p | \|\boldsymbol{x}\| \leq a\}$ containing $\mathcal{X}$ s.t.

$$\frac{\kappa_\ell}{V(B_a)} \leq p_X(\boldsymbol{x}_t) \leq \frac{\kappa_u}{V(B_a)},$$

for every $\boldsymbol{x}_t \in \mathcal{X}$. (Notice that this is the same a used in the F_p).

Following the same steps as in Wang *et al.* (2009), but noting that the assumptions concerning the distribution of $\boldsymbol{X}$ are now comparatively more relaxed in the sense that its support no longer needs to coincide with the closed ball B_a, one can show that the transformed random variable $U_{\boldsymbol{\beta}} \equiv F_p(X\boldsymbol{\beta})$ is quasi-uniformly distributed on $[0,1]$ (see Appendix A as well) in the sense that its density function is completely bounded, i.e. $0 < \kappa_\ell \leq p(u_{\boldsymbol{\beta},t}) \leq \kappa_u < \infty$, where $u_{\boldsymbol{\beta},t} \equiv F_p(\boldsymbol{x}_t'\boldsymbol{\beta})$. Therefore, it is reasonable to use equally spaced knots over the unit interval $[0,1]$, corresponding to the B-splines base. It should be noted yet that there is another possibility that should not be ruled out. Namely, the direct estimation of g from the data without the intermediation of F_p. However, this would turn out the choice of an appropriate set of the knots corresponding to the B-splines considerably more difficult, in the sense that they should be chosen, in this case, essentially based on the observed data. Unfortunately, such kinds of algorithms and strategies are not in the scope of this article.

The link function G is said to be a *spline* if it is a piecewise polynomial function of some arbitrary order k and some degree of smoothness which is used to guarantee continuity and differentiability up to a pre-determined order. Informally, as outlined in Eilers *et al.* (1996), a B-spline, B, of order k is a function with compact support consisting of $k+1$ polynomial components of order k continuously connected in the inner knots. These features are very interesting since they strongly simplify analysis and computational implementation, but for more specific details on the construction of a spline basis, we refer the reader to de Boor (2001). To estimate G, we write it as a linear combination of B-splines, $G(u) = \sum_{i=1}^M a_i B_i(u)$, where M is the number of elements in the basis and B_i is its ith member. Hence $\boldsymbol{y} = B(\boldsymbol{\beta})\boldsymbol{a} + \delta\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_T)'$ and $B(\boldsymbol{\beta}) = (B_j(u_{\boldsymbol{\beta},i}))_{i,j}$, with $i = 1, ..., T$ and $j = 1, ..., M$. In particular, by conditioning $\boldsymbol{\epsilon}$ on the latent variables σ_t, it follows that $\boldsymbol{y}|\boldsymbol{\sigma}; \boldsymbol{a}, \boldsymbol{\beta}, \delta^2 \sim \mathcal{N}\left(B(\boldsymbol{\beta})\boldsymbol{a}, \delta^2 W\right)$, with $W = \mathrm{diag}(\sigma_1^{-1}, ..., \sigma_T^{-1})$.

[1] F_p is the c.d.f. of the r.v. $a(2V-1)$, where V is Beta$((p+1)/2, (p+1)/2)$.

The larger the number of knots, the more the estimated function tends to interpolate the observed data, resulting then in a high variance estimate. To avoid this, Eilers *et al.* (1996) suggests penalizations on the adjacent splines coefficients (P-splines),

$$\lambda \sum_{j=d+1}^{M} (\Delta^d a_j)^2, \tag{3}$$

where λ plays the role of a smoothing parameter and Δ is the difference operator defined by $\Delta a_j = a_j - a_{j-1}$. This approach is computationally very attractive when compared to other kinds of penalization methods such as smoothing splines and it can be briefly represented by using the matrix form of the operator difference, denoted here by $\tilde{K}_d$ or just $\tilde{K}$ if no reference to d is necessary.

Under the Bayesian approach, penalties as in (3) are replaced by their stochastic counterparties. The most natural choice would be to write them as random walks as, for example, a first order random walk in the case of first differences, and second-order random walks in the case of second differences, $a_j = a_{j-1} + u_j$ or $a_j = 2a_{j-1} - a_{j-2} + u_j$, where $\{u_j\}$ are white noise. Lang *et al.* (2004) assume that $u_j|\tau^2 \sim \mathcal{N}(0, \tau^2)$ and that a_1, or a_1 and a_2, have noninformative priors. It is interesting to notice that τ^2 works as the inverse of the smoothing parameter λ as used in the frequentist case. However, a consequence of the above definition is that

$$p(\boldsymbol{a}|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{a}'\tilde{K}\boldsymbol{a}\right), \tag{4}$$

and since $\tilde{K}$ is singular (rank $\tilde{K} = M - 1$), the prior (4) is improper and, keeping $\tilde{K}$ this way, it would force us to impose conditions on the splines and data to guarantee a proper posterior. Instead we chose to adapt (4) by suggesting the modified version

$$p(\boldsymbol{a}|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{a}'(\tilde{K} + \lambda_a I)\boldsymbol{a}\right), \tag{5}$$

where I is a $M \times M$ identity matrix and $\lambda_a > 0$ is a (small) tuning parameter, inspired on ridge regressions, in order to make the prior distribution proper. In order to get some intuition behind the parameter λ_a we notice that (5), as well as (4), imposes $\boldsymbol{a}$ to be priorly distributed as a Normal random variable, both with the same covariance structures, but with different variances (i.e. larger in the modified version). In practical terms, this means assigning a comparatively less informative priori to $\boldsymbol{a}$, but essentially maintaining the same statistical relationships (correlations) among its elements. In other words, the inclusion of λ_a moves the distribution of $\boldsymbol{a}$ from a proper subspace of the M-dimensional Euclidean space to the entire space.

Another important issue to be considered is the one about the selection of λ_a and, of course, the underlying error distribution. There are many ways to address this problem and perhaps the most well known methods are the cross-validation or generalized cross-validation as in Hastie *et al.* (2001). However, there are other examples that could be included in such a list and with, say, a Bayesian flavor. In fact, we could consider (i) the proposal in Park *et al.* (2008) for the Bayesian Lasso, which is based on the use of a Monte Carlo EM algorithm to provide maximum likelihood estimates of the smoothing parameter at each iteration of the algorithm, or (ii) the log-pseudo marginal likelihood (LPML) as suggested by Geisser *et al.* (1979). Another approach, connected to the idea of cross-validation, would be the use the of the leave-one-out cross-validation using the simulated values of the parameters in the MCMC algorithm or even the smoothed leave-one-out cross-validation as discussed in Vehtari *et al.* (2017), which is just a way of properly approximating the predictive density when one data point is removed from the sample, i.e. $p(y_t|\boldsymbol{y}_{(-t)})$. In this paper, λ_a was chosen empirically by comparing metrics based on DIC, cross-validation and out-of-sample values. It is worth noting, however, that, as our experience indicates, small values of λ_a (~ 0.01) work generally equally well. On the other hand, larger values of λ_a ($\gg 1$) tend to excessively smooth the target function estimate. Finally, as a last remark, we will agree that to keep notation simple, we will denote from now on $\tilde{K} + \lambda_a I_d$ just by K.

As a last remark in this section, notice that conditioning $\boldsymbol{y}$ on the latent random vector $\boldsymbol{\sigma}$, we have

$$\log p(\boldsymbol{a}, \boldsymbol{\beta}, \delta^2, \tau^2; \boldsymbol{\sigma}, \boldsymbol{\zeta}|\boldsymbol{y}) \cong \frac{1}{2\delta^2}(\boldsymbol{y} - B(\boldsymbol{\beta})\boldsymbol{a})' W^{-1}(\boldsymbol{y} - B(\boldsymbol{\beta})\boldsymbol{a}) - \frac{1}{2\tau^2}\boldsymbol{a}' K \boldsymbol{a}, \qquad (6)$$

where $\cong$ stands for equality up to an additive constant (i.e. all additive terms unrelated to $\boldsymbol{a}$ in the right side of (6) were omitted for simplicity). Hence the posterior maximum likelihood estimate of $\boldsymbol{a}$ corresponds to a penalized weighted least-squares estimate.

4. Likelihood and Priors

We have already set up priors for the B-splines coefficients $\boldsymbol{a}$, so that $\boldsymbol{a}|\tau^2 \sim \mathcal{N}(\mathbf{0}, \tau^2 K^{-1})$. Now we do the same to the other priors as follows.

4.1 Likelihood and Modeling by Using Latent Variables

Modeling data as in (2), we get $\boldsymbol{y}|\boldsymbol{\sigma} \sim \mathcal{N}\left(B(\boldsymbol{\beta})\boldsymbol{a}; \delta^2 W\right)$, where $\sigma_t \overset{\text{iid}}{\sim} h$ and $W = \left(\sigma_1^{-1}, ..., \sigma_T^{-1}\right)$. Of course, the above random variables are conditioned on the transformed inputs $u_{\boldsymbol{\beta},t}$. The random variables σ_t work as weights which neutralize the effects of extreme values or outliers on the estimate. Besides, such hierarchical structure, which is a consequence of the use of the latent variables σ_t, turns the data analysis much simpler.

4.2 Priors

- **Index Vector ($\boldsymbol{\beta}$):** For $\boldsymbol{\beta}$, we assume a von Mises-Fisher distribution with concentration parameter κ and mean direction $\boldsymbol{\beta}_{prior}$ as prior:

$$p(\boldsymbol{\beta}|\boldsymbol{\beta}_{prior}, \kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)} \exp\{\kappa \boldsymbol{\beta}'_{prior}\boldsymbol{\beta}\},$$

  where $I_{p/2-1}$ denotes the Bessel function of first kind.

- **Scale and Smoothing Parameters (δ^2 and τ^2):** We assume an Inverse-Gamma prior distribution for the scale and smoothing parameters. More precisely, we set $\delta^2 \sim \text{IG}(\alpha_0, \gamma_0)$ and $\tau^2 \sim \text{IG}(\alpha_1, \gamma_1)$, so that

$$p(\delta^2|\alpha_0, \gamma_0) = \frac{\gamma_0^{\alpha_0}}{\Gamma(\alpha_0)}(\delta^2)^{-\alpha_0-1} \exp\left\{-\frac{\gamma_0}{\delta^2}\right\}$$

and

$$p(\tau^2|\alpha_1, \gamma_1) = \frac{\gamma_1^{\alpha_1}}{\Gamma(\alpha_1)} \exp\left\{-\frac{\gamma_1}{\tau^2}\right\},$$

where $\alpha_i, \gamma_i > 0$, for $i = 0, 1$.

- **Mixture Distribution Parameters and Hyperparameters ($\boldsymbol{\zeta}$):** The prior distribution of $\boldsymbol{\zeta}$ must be set up on a case-by-case basis since it obviously depends on the mixture distribution $h_{\boldsymbol{\zeta}}$. If there is no such parameter or if it is already known, we may leave this prior aside and proceed the analysis without it.

5. Sampling Scheme

We sample the parameters from their joint posterior distribution via the Metropolis-within-Gibbs algorithm, since it is not possible to represent the full conditional distributions for all parameters as well as the full conditional distribution for the latent variables σ_t in terms of standard distributions. We assume that $\boldsymbol{\zeta}$ is known, but later we will see the case where the degrees of freedom ($\zeta = \nu$) of the Student's t distribution is unknown. One can find the derivation of the full conditionals in Appendix.

- **B-splines coefficients ($\boldsymbol{a}$):** $\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\sigma}; \boldsymbol{\beta}, \delta^2, \tau^2 \sim \mathcal{N}_p(\mathbf{m}; \Delta)$, with

$$\Delta \equiv \left(\frac{1}{\delta^2} B(\boldsymbol{x}; \boldsymbol{\beta})' W^{-1} B(\boldsymbol{x}; \boldsymbol{\beta}) + \frac{1}{\tau^2} K_d\right)^{-1} \quad \text{and} \quad \mathbf{m} \equiv \frac{1}{\delta^2} \Delta B(\boldsymbol{x}; \boldsymbol{\beta})' W^{-1} \boldsymbol{y}.$$

- **Scale parameter (δ^2):** $\delta^2|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\sigma}; \boldsymbol{a}, \boldsymbol{\beta}, \overline{\tau} \sim \text{IG}\left(\alpha_0 + \frac{T}{2}, \gamma_0 + \frac{1}{2}\mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a})' W^{-1} \mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a})\right)$, and $\mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a}) \equiv \boldsymbol{y} - B(\boldsymbol{x}; \boldsymbol{\beta})\boldsymbol{a}$ is defined as the vector of residuals given the parameters $\boldsymbol{\beta}$ and $\boldsymbol{a}$.
- **Smoothing parameter (τ^2):** $\tau^2|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\sigma}; \boldsymbol{a}, \boldsymbol{\beta}, \delta^2 \sim \text{IG}\left(\alpha_1 + \frac{M}{2}, \gamma_1 + \frac{1}{2}\boldsymbol{a}' K \boldsymbol{a}\right)$.
- **Latent vector (σ):** the former full conditional distributions were explicitly derived, but this is not possible for $\boldsymbol{\sigma}$, $\boldsymbol{\beta}$ and, when appropriate, $\boldsymbol{\zeta}$. For $\boldsymbol{\sigma}$, we have

$$\begin{aligned} p(\boldsymbol{\sigma}|\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{a}, \boldsymbol{\beta}, \delta^2, \tau^2) &\propto \prod_{t=1}^{T} p(y_t|\boldsymbol{x}, \sigma_t; \boldsymbol{a}, \boldsymbol{\beta}, \delta^2) h(\sigma_t|\boldsymbol{\zeta}) \\ &\propto \prod_{t=1}^{T} \sigma_t^{1/2} \exp\left\{-\frac{r_t(\boldsymbol{\beta}; \boldsymbol{a})^2}{2\delta^2}\sigma_t\right\} h(\sigma_t|\boldsymbol{\zeta}), \end{aligned}$$

where $r_t(\boldsymbol{\beta}; \boldsymbol{a})$ is the tth component of $\mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a})$, and it is possible to reduce dimensionality by sampling one σ_t at a time for $t = 1, ..., T$. However, depending on $h(\cdot|\boldsymbol{\zeta})$, it is not possible to represent the posterior full conditional distribution of $\boldsymbol{\sigma}$ as some standard distribution. To overcome this difficulty, a Metropolis-Hastings step is introduced into the Gibbs sampler. Of course, h could be used as a prior proposal ($\sigma_t^* \sim h(\sigma_t|\boldsymbol{\zeta})$). However, one must be aware that if the likelihood p and the prior (and proposal) h are not concentrated over the same region, this method would be inefficient, and a more sophisticated approach would be necessary. On the other hand, fortunately, there are some interesting cases for which the above posterior distribution may be written explicitly (see Fernandez *et al.* (2000) and the Appendix).

- **Linear component (β):** the (posterior) full conditional distribution is given by

$$p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\sigma};\boldsymbol{a},\delta^2,\tau^2) \propto p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\sigma};\boldsymbol{a},\boldsymbol{\beta},\overline{\delta})\mathbb{I}(\boldsymbol{\beta}'\boldsymbol{\beta}=1 \text{ and } \beta_1>0),$$

and since this distribution does not match up with any standard distribution, we use a Metropolis-Hastings step to sample from it. The von Mises-Fisher distribution, with mean direction $\boldsymbol{\beta}_0$ and concentration parameter $\kappa > 0$, is taken as proposal, so that, given $\boldsymbol{\beta}_0$ (the vector sampled in the previous step of the algorithm),

$$\boldsymbol{\beta} = \begin{cases} \boldsymbol{\beta}_0, & \text{with probability } 1-\rho(\boldsymbol{\beta}_0,\boldsymbol{\beta}^*), \\ \boldsymbol{\beta}_*, & \text{with probability } \rho(\boldsymbol{\beta}_0,\boldsymbol{\beta}^*), \end{cases}$$

where

$$\rho(\boldsymbol{\beta}_0,\boldsymbol{\beta}_*) = \min\left\{\exp\left\{-\frac{1}{2\delta^2}(\mathbf{r}_*-\mathbf{r}_0)'W^{-1}(\mathbf{r}_*+\mathbf{r}_0)\right\},1\right\},$$

is the acceptance probability with $\mathbf{r}_* \equiv \mathbf{r}(\boldsymbol{\beta}_*,\boldsymbol{a})$ and $\mathbf{r}_0 \equiv \mathbf{r}(\boldsymbol{\beta}_0,\boldsymbol{a})$.

To conclude the section, we would like to note that, in relation to the sensitivity of the prior distributions as considered above with respect to their hyperparameters, we have tried, along the data analysis, for both simulations and application, several different configurations for such hyperparameters, and, from our experience, we have noticed that data fitting is quite robust with respect to these choices in the sense that the sampled values (through the MCMC algorithm) tend in any case to oscillate around the real parameter values (as in the simulation cases), and around very similar values (as in the case of the application).

5.1 A Note on Sampling the Degrees of Freedom for Student's t Errors

There are some alternatives to sample the degrees of freedom of the Student's t distribution when they are unknown (for further details, we refer the reader to Geweke (1993), the references therein, and to Fonseca *et al.* (2008)). However, since the specific analysis of the Student's t case is not the focus of this article, we have decided to follow the specifications in Cabral *et al.* (2012) and Bandyopadhyay *et al.* (2015) and to assume a hierarchical structure for the prior distribution of ν. More precisely, we assume that $\nu|\lambda$ follows a (left) Truncated Exponential distribution at ν_0 with parameter λ, so that $\nu - \nu_0 \sim \text{Exponential}(\lambda)$ $(\nu \geq \nu_0 \geq 0)$, and, as for parameter λ, it is assumed to follow an Uniform distribution on the interval $(\lambda_\ell, \lambda_u)$. In particular, it should be noticed that a suitable choice of ν_0 (e.g. $\nu_0 = 2$) guarantees the suitable property of finite variance for the errors associated with the model. Regarding the hyperparameters λ_ℓ and λ_u, they can be chosen by imposing bounds on the moments of ν. For example, Cabral *et al.* (2012) fixed $\lambda_\ell = 0.02$ and $\lambda_u = 0.5$, so that the expected value of ν is in the interval $[2, 50]$ — the same bounds adopted by Bandyopadhyay *et al.* (2015). The derivations of the expressions below can be found in Appendix C. The posterior full conditional is given by

$$p(\nu|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\sigma};\boldsymbol{a},\boldsymbol{\beta},\delta^2,\tau^2) \propto \frac{\left(\frac{\nu}{2}\right)^{\frac{T\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)^T}\exp\left\{-\nu\left(\frac{1}{2}\sum_{t=1}^{T}(\sigma_t-\log\sigma_t)+\lambda\right)\right\}\mathbb{I}(\nu>\nu_0). \quad (7)$$

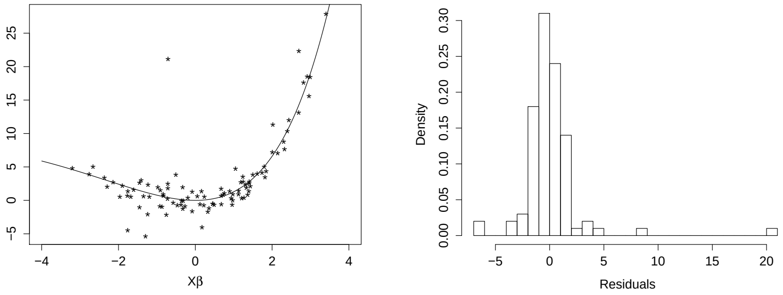
Metropolis-Hastings is used to sample ν from (7) and the proposal distribution was chosen to be the exponential distribution with parameter λ_* (to be calibrated on a case-by-case basis).

Just as a final remark, since Metropolis-Hastings was already used to sample $\boldsymbol{\beta}$, these two similar steps inside the Gibbs algorithm could be reduced to a single one as suggested by Müller (1993), with the effect of reducing the rejection rate of the proposals. This would produce a global approximation for the posterior full conditional of $\boldsymbol{\beta}$ and ν, instead of local approximations. Derivations are very similar to those in Appendix C and, for the sake of brevity, shall be omitted.

6. Examples

6.1 Simulation Study

For the sake of comparison, this simulation study follows almost the same setup proposed by Antoniadis *et al.* (2004). A sample of size 100 was generated according to the model defined by the link function $g(u) = u^2 e^{u/4}$, vector index $\beta = (2, 1, 1, 1)'/\sqrt{7} = (0.756; 0.378; 0.378; 0.378)'$ and predictors $\boldsymbol{x}_t$ independently and uniformly generated from the hypercube $[-3, 3]^4$. The link function was approximated by cubic B-splines with the smoothing matrix $\widetilde{K}$ determined by taking second differences and M chosen empirically by assessing the estimation performance for several different values by means of metrics such as DIC, deviance and mean squared error. Some differences with Antoniadis *et al.* (2004) are inevitable due to the distinct data sets and some idiosyncrasies of the estimation methods. In fact, in Antoniadis *et al.*, the noise is assumed to be Normal, while we assumed a heavier tailed Student's t distribution with 2.2 degrees of freedom and scale parameter $\delta^2 = 1$. Moreover, we use P-splines with smoothing order $(d = 2)$, while in Antoniadis *et al.* (2004), the penalization is directly over $\|\boldsymbol{a}\|^2$. The dispersion of the data set around the link function and the histogram of the residuals under the severe scenario can be seen in Figure 1. Such severity, in particular, can be noticed in Figure 1(a) by the presence of a very extreme value, which is in fact caused by the noise, whose distribution is extremely heavy tailed.



(a) Link Function

(b) Residuals Histogram

Figure 1. 1(a): Link function and the simulated data under Student's t errors with $\nu = 2.2$. 1(b) Histogram of the theoretical residuals (i.e. calculated with the true parameters) of the simulated model.

As an illustration, Table 1 shows the effect of M over the model estimate. Increasing M, makes the link function estimate tend to interpolate data and so it is expected that the deviance will also be greater. The average deviance on the other hand is expected to decrease as the number of parameters increase, and so the combination of both (the deviance information criterion, or DIC) would lead us to large values of M. However these values also imply greater complexity (average deviance minus deviance), so a compromise between these values suggest an intermediate value of $M = 65$. This choice is clearly supported by the posterior MSE of the predicted responses and $\boldsymbol{\beta}$.

Table 1. Model assessment by the number of elements in the splines basis (M). Mean square error, point estimate deviance, average deviance and DIC ($\nu = 3$).

M	30	40	50	60	70	80	90	100	200
MSE ($\boldsymbol{\beta}$)	0.79	0.68	0.64	0.52	0.45	0.53	0.59	0.69	0.82
MSE	11.10	6.53	5.89	2.88	3.27	3.36	8.99	10.29	45.17
Deviance	680.45	709.24	732.59	743.84	760.04	764.70	772.90	770.92	868.32
Avg Dev	595.01	574.20	547.05	528.39	508.38	495.31	469.24	464.92	438.84
DIC	509.57	439.17	361.52	312.94	256.72	225.92	165.57	158.92	9.36

Regarding the proposal distribution for $\boldsymbol{\beta}$, we accept the value suggested in Antoniadis *et al.* (2004) and set up the concentration parameter κ equal to 1000. As in the aforementioned article, we estimated the parameters of interest by taking the posterior means of the 8000 generated values after the burn-in period of size 2000. We set the prior parameters for the scale δ^2 as $\gamma_0 = \alpha_0 = 400$ and for τ^2, we set $\alpha_1 = 100$ and $\gamma_1 = 8000$.

To estimate the degrees of freedom related to the error distribution, we considered several values of ν and the results can be seen in Table 2. Based on the same rationale as before, we decided to take $\nu = 2.5$ which is not far from the actual value of 2.2. Moreover, by considering subsamples of the same data set of sizes $T' = 25$ and $T' = 50$ (Tables 3 and 4) we see as expected that an appropriate choice for the error distribution becomes significantly more important for smaller data sets.

Table 2. Model assessment by degrees of freedom (ν). Mean square error, point estimate deviance, average deviance and DIC ($M = 65$).

ν	∞	2	3	4	5	6	7
MSE ($\boldsymbol{\beta}$)	0.64	0.60	0.49	0.65	0.54	0.72	1.19
MSE	3.19	2.75	1.59	2.36	3.37	7.59	10.01
Dev	2,885.63	657.08	733.39	751.54	809.74	848.27	891.82
Avg Dev	670.53	545.96	511.45	583.17	604.17	617.99	624.66
DIC	-1,544.58	434.84	289.52	414.79	398.61	387.71	357.50

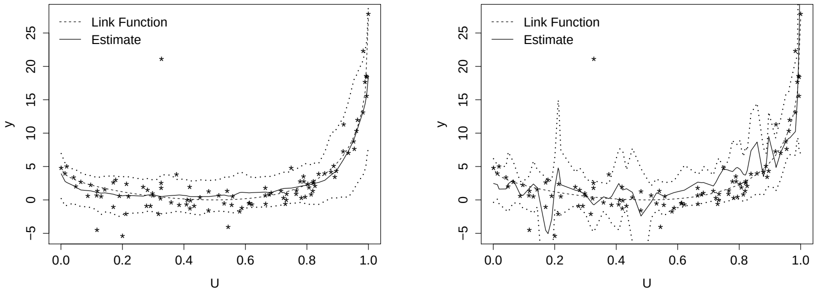
Table 3. Model assessment by degrees of freedom (ν) with 25 data points. Mean square error, point estimate deviance, average deviance and DIC ($M = 65$).

ν	∞	2	3	4	5	6	7
MSE ($\boldsymbol{\beta}$)	0.52	0.70	0.70	0.69	0.68	0.64	0.58
MSE	10.98	7.18	8.15	10.87	8.67	8.50	6.20
Dev	830.83	189.83	200.05	225.23	236.16	237.29	240.87
Avg Dev	171.41	134.00	133.67	137.45	137.92	136.85	142.87
DIC	-488.01	78.19	67.29	49.67	39.68	36.42	44.88

Table 4. Model assessment by degrees of freedom (ν) with 50 data points. Mean square error, point estimate deviance, average deviance and DIC ($M = 65$).

ν	∞	2	3	4	5	6	7
MSE ($\boldsymbol{\beta}$)	0.75	0.57	0.65	0.42	0.50	0.58	0.54
MSE	16.24	5.54	8.75	19.53	7.51	7.78	7.11
Dev	1,393.01	342.82	380.00	458.47	465.77	479.72	496.33
Avg Dev	416.28	223.96	230.79	232.29	238.77	238.84	256.69
DIC	25.50	105.10	81.58	6.11	39.68	-2.05	17.04

The graphs in Figure 2 illustrate (i) how the observed data are sparse (because of the heavy tails of the distribution attributed to noise), and (ii) how well the estimated curve fits the data and thus approximates quite satisfactorily the link function. Notice that



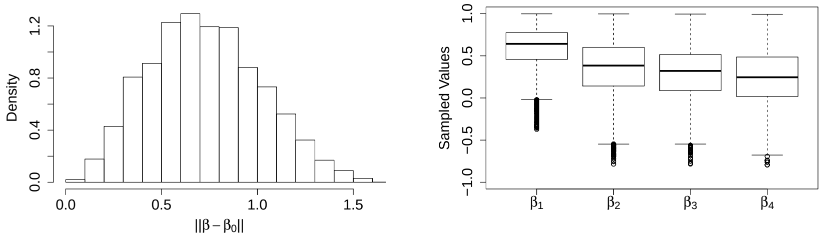
(a) Student's t

(b) Normal

Figure 2. Estimates of the link function the assumption of (a) Student's t and (b) Normal errors. Dotted lines refer to 95% confidence bands based on the posterior distribution and estimated from the MCMC simulated values.

these are the estimates of the transformed link function G. It is evident that the proposed method offers the possibility of considerably more robust estimates which can be used in the presence of extreme values.

Regarding the index vector, we notice that its posterior mean, $\widehat{\boldsymbol{\beta}} = (0.731; 0.458; 0.360; 0.356)'$ is quite close to the actual value. Figures 3 and 4 show how well they fit the real parameter and how they are distributed *a posteriori*. They also



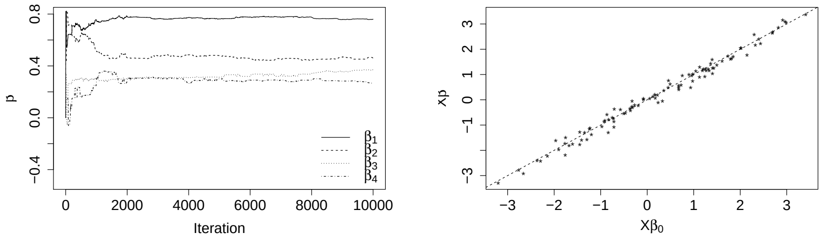
(a) $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|$ histogram

(b) Linear parameters boxplots

Figure 3. 3(a): Histogram based on the posterior distribution of $\|\boldsymbol{\beta} - \beta_0\|$, where $\|\boldsymbol{\beta}_0\|$ are the real index vector. 3(b) Boxplot based on the posterior distribution of $\boldsymbol{\beta}$.

indicate that the generated Markov chain for the index vector converges quickly to its stationary distribution. It is shown that the estimated design points, $X\widehat{\boldsymbol{\beta}}$, fits well the initially proposed design, i.e, $X\boldsymbol{\beta}$. The dashed line in 4(b) is simply the diagonal line $y = x$. Finally, the simulations of the smoothing parameter $(1/\tau)$ and scale parameter (δ) can be seen in Figure 5. The posterior means (± std. dev.) are 0.1184 ± 0.0056 and 1.031 ± 0.036, respectively.

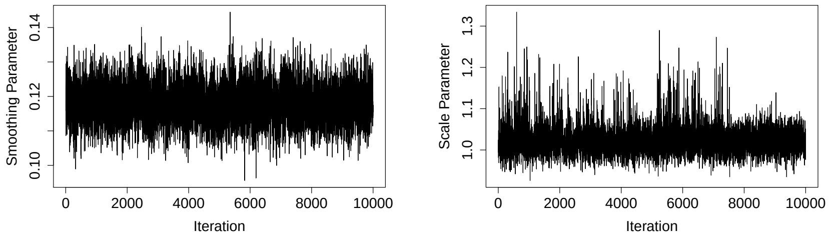
The point in using heavy-tailed distributions is that they respond well to the presence of extreme or influential values. For example, Castro *et al.* (2014) consider the use of splines and Scale Mixtures of Normals to analyze censored partially linear models and show how to perform case-deletion influence diagnostics in the presence of heavy tailed distributed errors. In our case, the suggested method also allows us to take advantage of considering heavy tailed distributions to the model error component by using the sampled values of the weights σ_t to identify extreme and influential points. For the sake of illustration, Figure 6(a) shows the graphs of the actual target function versus the estimated function. In addition to them, we also see the values of the response variable in terms of $X\boldsymbol{\beta}$ over the cited curves. On the other hand, Figure 6(b) shows the weights σ_t also in terms of $X\boldsymbol{\beta}$. It is interesting to note that they capture the outliers located to the left of the value



(a) $\boldsymbol{\beta}$ samplings

(b) $X\boldsymbol{\beta}$ *vs.* $X\widehat{\boldsymbol{\beta}}$ scatterplot

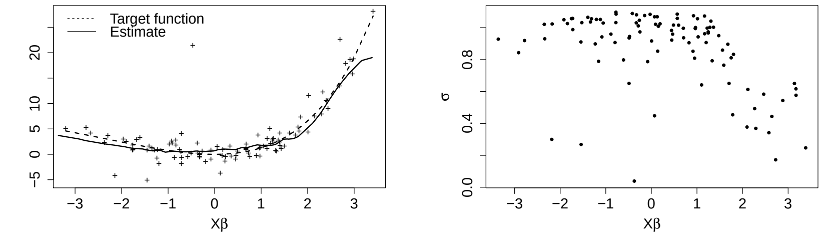
Figure 4. 4(a): Convergence of the Index Vector simulated components. 4(b): Estimated transformed design points, $X\boldsymbol{\beta}$ (horizontal axis), *vs.* the actual transformed design points, $X\widehat{\boldsymbol{\beta}}$ (vertical axis). The solid line is just least squares linear fit (with intercept 0.01 and slope 0.97).



(a) Smoothing Parameter

(b) Scale Parameter

Figure 5. 5(a): MCMC simulated values for $1/\tau$. 5(b): MCMC simulated values for δ.



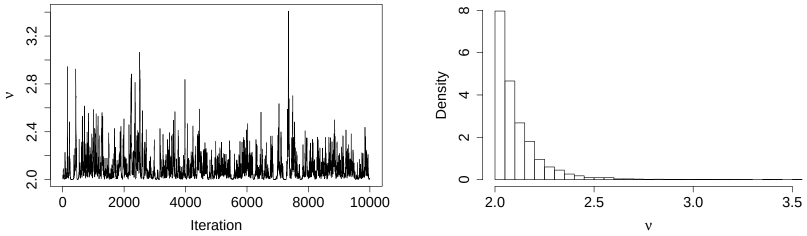
(a) Target function *vs.* estimate

(b) σ weights *vs.* $X\boldsymbol{\beta}$

Figure 6. Comparison between the (real) target function and its estimate and the corresponding weights of each observation obtained from the latent variables σ_t.

$\boldsymbol{x}'\boldsymbol{\beta} = 0$. The values at the right end of the chart also tend to receive a lower weighting, and this is due to the rapid growth of the target function in this region of the graph.

Finally, we adjust the data assuming that the errors follow Student's t distribution, but without setting the degrees of freedom ν in advance. The prior distribution corresponding to ν as well the associated hierarchical structure is as described in Section 5.1. In Figure 7, we can see the samplings of ν drawn from its posterior distribution, as well as its histogram, and the results are consistent with the previous one. Moreover, from the posterior samplings we get the 95% credible interval $(2.01; 2.43)$ and a point estimate, the posterior mean, equal to 2.11. To conclude, by way of comparison, we also report the posterior mean of $\boldsymbol{\beta}$ as being equal to $(0.703; 0.275; 0.516; 0.405)'$, which remains very close to the real value.



(a) Degrees of freedom samplings (b) Degrees of freedom histogram

Figure 7. Samplings and histogram associated to posterior distribution of Student's t degrees of freedom (ν) obtained through the proposed MCMC algorithm.

6.2 An Application

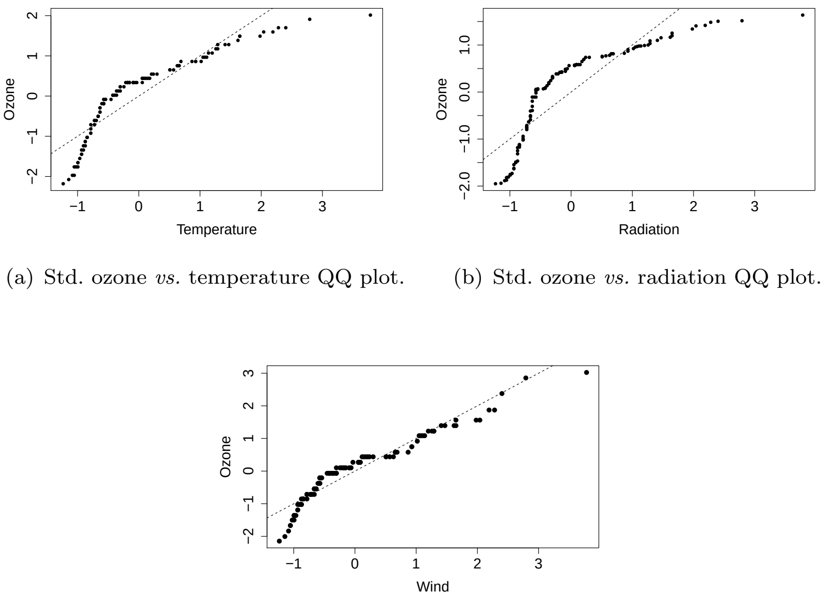
As an illustration, we apply the proposed model to a real data set with 111 daily measurements (from May to September, 1973) exploring the relationship between environmental variables: concentration of the air pollutant ozone (output variable) and three meteorological indicators: solar radiation, windspeed and daily maximum temperature (input variables). These observations were taken in New York and measured in parts per billion, Langleys, miles/hour and degrees Fahrenheit, respectively. This data set is available in the R package *ElemStatLearn*.

For comparison, this same data set was already studied by using the single-index model, see Park *et al.* (2005) and Yu *et al.* (2002). In the first one, the link function was approximated by using wavelets and the model error was assumed to follow a Normal distribution. This assumption makes the link function estimate to be less smooth than ours, since it is more sensitive to extreme values. Besides that, in Yu *et al.* (2002), five different models were compared, including a linear model and some multivariate semiparametric models, and the results indicate that the single-index and additive models are the best suited to fit the data. In particular, in favor of using heavy-tailed distributions (and many of which can be represented as scale mixtures of Normals), the QQ-plots in Figure 8, constructed using the standardized variables derived from Ozone (output) and Temperature, Radiation and Wind (inputs) indicate a deviation from the Standard Normal distribution toward more leptokurtic distributions — as pointed out by the considerable spacing between the QQ-plot dots and the dashed line. Looking only at the output variable alone, Figure 9 also indicates a heavy-tailed behavior, as one can see through the QQ-plot of the empirical distribution of the standardized variable against the Normal distribution, or the histogram of the excess kurtosis calculated using a Bootstrap sample of size 5000, whose estimate is 1.13 with a 95% confidence interval given by $(-0.56; 3.16)$. Finally, it should be said that this phenomenon is further reinforced by the scatter plot of the output (Ozone) in relation to the transformed variables as indicated in Figure 10, as well as the boxplots in Figure 11 in the sense that estimates are much more volatile when using the Normal distribution as the error distribution.

In analyzing data we considered four different distributions: the Student's t (for some different degrees of freedom), the Double Exponential[1] with parameters $\mu = 0$ and $b = 1$, the Standard Slash and the Normal distributions. For comparison, we display in Table 5 the respective deviations and average deviations.

Tables 5 and 6 also suggest that it is not necessary to use large values of M to get a good

[1] The Double Exponential distribution on the real line is defined by the pdf $p(x) = \exp(-|x - \mu|/b)/(2b)$, where $\mu \in (-\infty, \infty)$ represents the location parameter (which coincides with its mean and median) and $b > 0$ is the scale parameter (with $2b^2$ corresponding to the variance of the distribution).

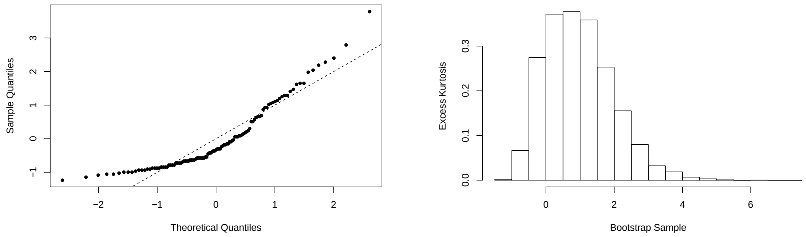


(a) Std. ozone *vs.* temperature QQ plot.

(b) Std. ozone *vs.* radiation QQ plot.

(c) Std. ozone *vs.* wind QQ plot.

Figure 8. Quantile-quantile plots of the standardized ozone level empirical distribution against the standardized inputs (temperature, ozone level and wind).



(a) Std. ozone *vs.* Std. Normal QQ plot.

(b) Excess kurtosis histogram.

Figure 9. Quantile-quantile plots of the (standardized) ozone level empirical distribution against the Normal and the histogram of the excess kurtosis obatained via Bootstrap resampling.

Table 5. Deviations and Average Deviations for different models and different values of M.

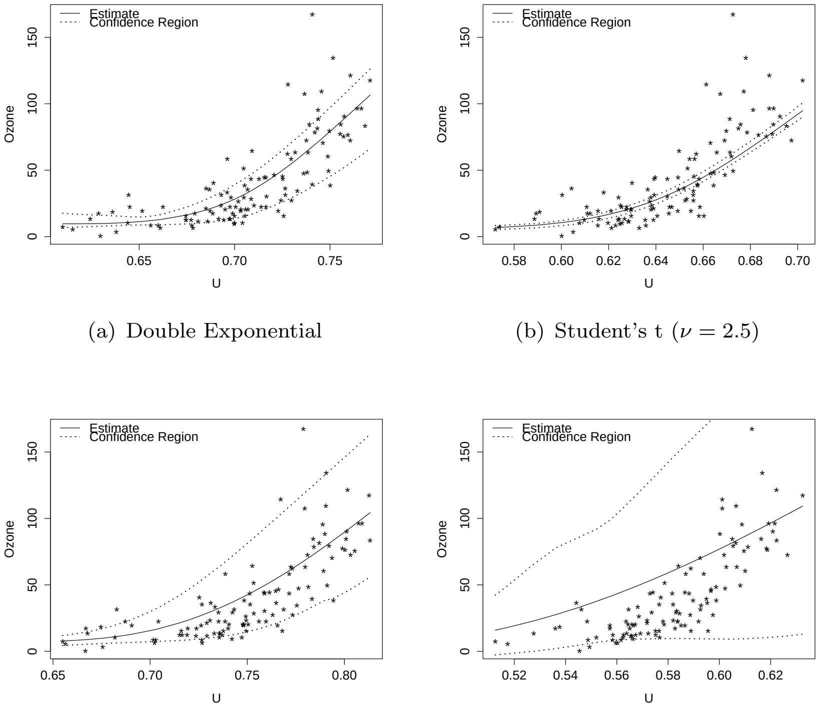
	M=15		M=20		M=25	
Distribution	Dev	Avg Dev	Dev	Avg Dev	Dev	Avg Dev
Normal	3,114.31	1,106.91	2,262.02	1,117.36	3,266.56	1,113.99
Student-t ($\nu = 2.5$)	2,127.94	1,531.03	2,141.52	1,489.19	2,105.18	1,557.55
Student-t ($\nu = 3$)	2,296.94	1,578.15	2,248.63	1,578.82	2.339,48	1,478.31
Student-t ($\nu = 4$)	3,617.11	1,869.02	2,647.86	1,630.34	2,805.28	1,724.12
Student-t ($\nu = 5$)	2,826.25	1,766.45	2,853.77	1,741.00	2,891.60	1,772.92
Student-t ($\nu = 6$)	3,002.71	1,872.66	3,089.53	1,796.27	3,092,92	1,817.01
Student-t ($\nu = 8$)	3,433.11	1,952.14	3,382.14	1,931.50	3,418.19	1,873.77
Double Exponential	1,358.37	2,038.61	1,735.26	1,749.11	1,746.60	1,734.53
Slash	2,091.76	1,262.74	2,258.88	1,201.14	2,152.54	1,158.87

fit and so we decided for $M = 20$. Besides, the numbers in the same table suggest that models allowing for fat tailed distributed errors fit better the link function to the data. Another clue is the fact that when the degrees of freedom associated to the Student's t distribution are not fixed *a priori*, the model tends to choose values close to 2. In particular, the Student's t and the Double Exponential distributions seem to be the best options. In

Table 6. Continuation of Table 5.

	M=30		M=35	
Distribution	Dev	Avg Dev	Dev	Avg Dev
Normal	6,725.41	1,116.87	11,682.11	1,118,14
Student-t ($\nu = 2.5$)	2,180.45	1,421.04	2,199.60	1,537.28
Student-t ($\nu = 3$)	2,317.68	1,494.56	2,231.08	1,740.25
Student-t ($\nu = 4$)	2,643.65	1,616.86	2,658.32	1,618.49
Student-t ($\nu = 5$)	2,821.61	1,760.72	2,949.00	1,665.47
Student-t ($\nu = 6$)	3,154.35	1,839.22	3,029.63	2,152.32
Student-t ($\nu = 8$)	3,426.39	1,847.13	3,482.59	2,300.99
Double Exponential	1,758.17	1,724.57	1,798.52	1,788.86
Slash	2,186.47	1,156.20	2,136.92	1,227.98

Figure 10, we compare the estimates of the link function Double Exponential, Student's t (with $\nu = 2.5$), Slash and Normal cases. The correspondent index vector estimates can be found in Table 7 and we can make at least three observations. First, we notice that the confidence bands for the Student's t and Double Exponential distributions are narrower, especially in the extremes as expected, since such distributions, because they have heavier tails than the Normal distribution, better accommodate extreme values and outliers. In fact, to accommodate extreme values under the assumption of a Normal distribution, the scale parameter has to be considerably larger than its counterparts associated with the more robust distributions (Student's, Double Exponential and Slash, for example) In broader confidence bands. Similar phenomenon occurs with the estimates of the parameters corresponding to the linear component of the model as discussed below. Secondly, the Gaussian model implies considerable higher levels of ozone in the atmosphere for lower values of $X\widehat{\boldsymbol{\beta}}$, which is not ratified when we the noise follows a distribution with heavier



(a) Double Exponential

(b) Student's t ($\nu = 2.5$)

(c) Slash

(d) Normal

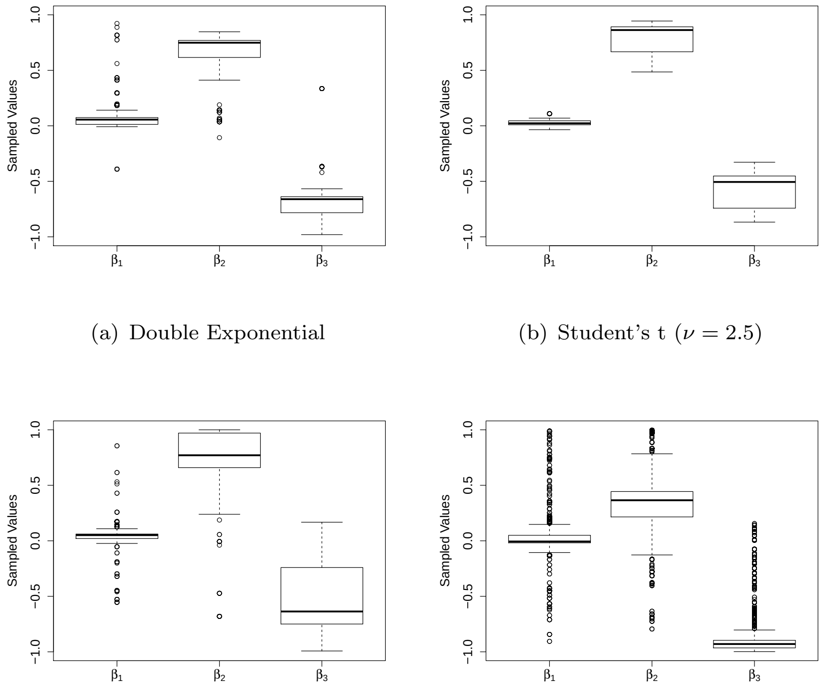
Figure 10. 10(a): Link function estimate and its 95% confidence region under the assumption of Double Exponential distributed errors. 10(b), 10(c) and 10(d): The same for the Student's t distribution ($\nu = 2.5$), Slash and Normal distributions, respectively.

tails. Moreover, it is clear that the Normal model is significantly more sensitive to extreme values which causes the link function to be more linear and, therefore, making the ozone concentration to increase at the same rate for any combination of climatic inputs. Finally, comparing how spread are the transformed values of $X\widehat{\boldsymbol{\beta}}$, as can seen in Figure 10, it is evident that the influence of input variables is not the same under the assumptions of normality and heavy tails.

Table 7. Posterior means and standard deviations of the climatic variables (Radiation, Temperature and Wind) and model parameters (Scale and Smoothing parameters).

	Coefficients				
Noise	Radiation	Temperature	Wind	Scale Parameter	Smoothing Parameter
Normal	0.02 (0.10)	0.36 (0.24)	-0.93 (0,18)	8.11 (1.22)	0.0984 (0.0079)
Student's	0.026 (0.025)	0.81 (0.12)	-0.58 (0.15)	1.161 (0.034)	0.0992 (0.0053)
Slash	0.048 (0.050)	0.84 (0.16)	-0.54 (0.31)	1.057 (0.028)	0.1025 (0.0059)
Double Exponential	0.051 (0.054)	0.70 (0.11)	-0.710 (0.098)	2.09 (0.17)	0.0979 (0.0058)

In Figure 11, one can see the boxplots corresponding to the estimates in Table 7. Again, it is clear that there are consequences in choosing a distribution with heavier tails or not. The Normal model tends to give somewhat less importance to Radiation and significantly less importance to Temperature when compared to the other models. On the hand, it emphasizes the weight of wind in the ozone concentration. In the Normal model, the index vector posterior distribution is clearly fat tailed, which makes its estimates have larger standard deviations.



(a) Double Exponential

(b) Student's t ($\nu = 2.5$)

(c) Slash

(d) Normal

Figure 11. 5(a): Boxplot for the sampled values of $\widehat{\boldsymbol{\beta}}$ under the Gaussian assumption. **Right pane:** the same under the Double Exponential assumption.

7. Extensions

Some possible improvements in the proposed models that will the subject of future works are: (a) estimation of curves with varying degree of smoothness and (b) model selection. In fact, we considered here only functions with a global amount of smoothing. This could be a drawback if the curvature of the target function varies strongly with the inputs. In Sheipl *et al.* (2009), this problem is dealt with by introducing a locally adaptive spline smoothing by using scale mixtures of Normals with locally varying exponential-gamma distributed variances for the differences of the P-spline. Introduction of such approach in our model would imply at least an extra level in the priors hierarchy and more complexity in the MCMC scheme. Regarding model selection, the open issue here is the selection of the parameter λ_a. As indicated in Section 3 and the references therein, there are ways of doing so using generic means available in the literature. However, it would be interesting to establish a specialized and fully Bayesian method for choosing the parameter λ_a (associated with the penalization matrix K). This is a topic that we are still exploring and which we intend to present in a future work. Another question of interest for us is to determine automated ways to select the mixture distribution h. As far as we know, there is no satisfactory way of doing that and inference must be done by testing the different possibilities individually and comparing them afterwards.

8. Conclusions

In this paper, we presented a Bayesian framework for estimation and inference of the parameters of interest, *i.e.* the parameters that make up the model (1). We presented here the formulas needed to implement each step of the MCMC algorithm used to sample the parameters according to their posterior distributions and the conditions under which the posterior distribution is proper. We emphasized the use of Student's t distribution, but we also showed how one could adapt the model to different choices of distributions for the noise. The computation, although computationally intensive, is straightforward to implement and has the advantage, when compared to Antoniadis *et al.* (2004), to associate a non informative prior to the index vector, besides being a completely Bayesian approach (with no regualarization).

The efficiency of the proposed methodology is confirmed through simulation study and application to real data. Concerning the application, the above methodology suggests that models based on normal distribution are inappropriate. Finally, we noticed that applications of such approach to more general models, such as models with varying degrees of smoothness and even multi-index models and model selection were left for future works.

Appendix A. On the Distribution of $U_{\boldsymbol{\beta}}$

We have included the subsequent argument just for convenience, for it follows essentially the same steps as described in Wang *et al.* (2009). It should be noted, however, that we are now applying them to a more relaxed set of hypotheses. That is, we do not require that the support of the pdf of $\boldsymbol{X}_t$ coincide with the closed ball B_a.

Let $X_{\boldsymbol{\beta}} = \boldsymbol{\beta}'\boldsymbol{X}$ and $p_{\boldsymbol{\beta}}$ be the pdf of $U_{\boldsymbol{\beta}}$. It follows then that $p_{\boldsymbol{\beta}}(u) = \left[F_p'(v)\right]^{-1} p_{X_{\boldsymbol{\beta}}}(v)$, where $v = F_p^{-1}(u)$ $(0 \leq u \leq 1)$. Define $D_v = \{x \in \mathrm{R}^p | v \leq x_{\boldsymbol{\beta}} \leq v + \Delta v\} \cap B_a$, so that

$P(v \leq X_{\boldsymbol{\beta}} \leq v + \Delta v) = P(\boldsymbol{X} \in D_v) = \int_{D_v} p_X(\boldsymbol{x}) d\boldsymbol{x}$. From Assumption 2,

$$\frac{\kappa_\ell V_p(D_v)}{V(B_a)} \leq P(v \leq X_{\boldsymbol{\beta}} \leq v + \Delta v) \leq \frac{\kappa_u V_p(D_v)}{V(B_a)},$$

where $V_p(D_v)$ and $V_p(B_a)$ are the p-dimensional volumes of D_v and B_a, respectively. On the other hand, $V_p(D_v) = V_{p-1}(I_v)\Delta v + o(\Delta v)$, where $I_v = \{x \in \mathrm{R}^p | x_{\boldsymbol{\beta}} = v\} \cap B_a$. Now, using the facts that

$$V_p(I_v) = \frac{\pi^{(p-1)/2}(a^2 - v^2)^{(p-1)/2}}{\Gamma((p+1)/2)} \quad \text{and} \quad V_p(B_a) = \frac{\pi^{p/2} a^p}{\Gamma(p/2 + 1)},$$

and the identity $\Gamma(z)\Gamma(z + 1/2) = 2^{1-2z}\pi^{1/2}\Gamma(2z)$, we have $V_p(D_v)/V_p(B_a) = F_p'(v)\Delta v + o(\Delta v)$. Hence, $\kappa_\ell \left(F_p'(v)\Delta v + o(\Delta v)\right) \leq P(v \leq X_{\boldsymbol{\beta}} \leq v + \Delta v) \leq \kappa_u \left(F_p'(v)\Delta v + o(\Delta v)\right)$, so that dividing the all termos in the above inequalities by Δv and doing $\Delta v \to 0$, we get $\kappa_\ell \leq F_p'(v)^{-1} p_{X_{\boldsymbol{\beta}}}(v) = p_{\boldsymbol{\beta}}(u) \leq \kappa_u$.

Appendix B. Full Conditionals

In the following, we will not make explicit, for simplicity, the dependence of the full conditional distributions on the specific parameters and hyperparameters associated to the scale factor, namely $\boldsymbol{\zeta}$.

- **B-splines coefficients:** writing $\Delta \equiv \left(\frac{1}{\delta^2} B(\boldsymbol{\beta})' W^{-1} B(\boldsymbol{\beta}) + \frac{1}{\tau^2} K_d\right)^{-1}$, it is clear that

$$p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{a}|\tau^2) \propto \exp\left\{-\frac{1}{2\delta^2}(\boldsymbol{y} - B(\boldsymbol{\beta})\boldsymbol{a})' W^{-1}(\boldsymbol{y} - B(\boldsymbol{\beta})\boldsymbol{a})\right\} \cdot \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{a}' K_d \boldsymbol{a}\right\}$$
$$\propto \exp\left\{-\frac{1}{2}\left((\boldsymbol{a} - \mathbf{m})' \Delta^{-1} (\boldsymbol{a} - \mathbf{m})\right)\right\},$$

where $\mathbf{m} \equiv \frac{1}{\delta^2} \Delta B(\boldsymbol{x}; \boldsymbol{\beta})' W^{-1} \boldsymbol{y}$. Now,

$$p(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\sigma}; \boldsymbol{\beta}, \delta^2, \tau^2) \propto p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\sigma}; \boldsymbol{a}, \boldsymbol{\beta}, \delta^2) p(\boldsymbol{a}|\tau^2)$$
$$\propto \exp\left\{-\frac{1}{2}\left((\boldsymbol{a} - \mathbf{m})' \Delta^{-1} (\boldsymbol{a} - \mathbf{m})\right)\right\},$$

so that $\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{u}_{\boldsymbol{\beta}}, \boldsymbol{\sigma}; \boldsymbol{\beta}, \delta^2, \tau^2 \sim \mathcal{N}_p(\mathbf{m}; \Delta)$.

- **Scale parameter:** let $\mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a}) \equiv \boldsymbol{y} - B(\boldsymbol{\beta})\boldsymbol{a}$ as the vector of residuals given the parameters $\boldsymbol{\beta}$ and $\boldsymbol{a}$, so

$$p(\delta^2|\boldsymbol{y}, \boldsymbol{u}_{\boldsymbol{\beta}}, \boldsymbol{\sigma}; \boldsymbol{a}, \boldsymbol{\beta}, \tau^2) \propto p(\boldsymbol{y}|\boldsymbol{u}_{\boldsymbol{\beta}}, \boldsymbol{\sigma}; \boldsymbol{a}, \boldsymbol{\beta}, \delta^2) p(\delta^2|\alpha_0, \gamma_0)$$
$$\propto \left(\frac{1}{\delta^2}\right)^{(2+\alpha_0+T/2)-1} \exp\left\{-\frac{1}{\delta^2}\left(\gamma_0 + \frac{1}{2}\mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a})' W^{-1} \mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a})\right)\right\},$$

so that

$$\delta^2|\boldsymbol{y}, \boldsymbol{u}_{\boldsymbol{\beta}}, \boldsymbol{\sigma}; \boldsymbol{a}, \boldsymbol{\beta}, \overline{\tau} \sim \mathrm{IG}\left(2 + \alpha_0 + \frac{T}{2}, \gamma_0 + \frac{1}{2}\mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a})' W^{-1} \mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a})\right).$$

- **Smoothing parameter:** analogously,

$$p(\tau^2|\boldsymbol{y},\boldsymbol{u}_{\boldsymbol{\beta}},\boldsymbol{\sigma};\boldsymbol{a},\boldsymbol{\beta},\delta^2) \propto p(\boldsymbol{a}|\tau^2)p(\tau^2|\alpha_1,\gamma_1) = \left(\frac{1}{\tau^2}\right)^{(2+\alpha_1+\frac{M}{2})-1} \exp\left\{-\frac{1}{\tau^2}\left(\gamma_1+\frac{\boldsymbol{a}'K\boldsymbol{a}}{2}\right)\right\},$$

so that

$$\tau^2|\boldsymbol{y},\boldsymbol{u}_{\boldsymbol{\beta}},\boldsymbol{\sigma};\boldsymbol{a},\boldsymbol{\beta},\delta^2 \sim \text{IG}\left(2+\alpha_1+\frac{M}{2},\gamma_1+\frac{\boldsymbol{a}'K\boldsymbol{a}}{2}\right).$$

- **Latent variable:** Starting with the Contaminated Normal distribution, we get $\sigma_t|y_t,\boldsymbol{u}_{\boldsymbol{\beta}};\boldsymbol{a},\boldsymbol{\beta},\delta^2,\tau^2 = 1$ with probability $1-\xi'$, or equal to λ^2 with probability ξ', where $\xi' = C_t\xi\lambda\exp\{-r_t(\boldsymbol{\beta};\boldsymbol{a})^2\lambda/(2\delta^2)\}$ and C_t is a normalizing constant which depends on the observation $(y_t;\boldsymbol{x}_t')'$. For the Student's t distribution, we have

$$p(\sigma_t|y_t,\boldsymbol{u}_{\boldsymbol{\beta}};\boldsymbol{a},\boldsymbol{\beta},\delta^2,\tau^2) \propto \sigma_t^{\frac{\nu+1}{2}-1}\exp\left\{-\frac{\nu+r_t(\boldsymbol{\beta};\boldsymbol{a})^2/\delta^2}{2}\sigma_t\right\},$$

so that

$$\sigma_t|y_t,\boldsymbol{u}_{\boldsymbol{\beta}};\boldsymbol{a},\boldsymbol{\beta},\delta^2,\tau^2 \sim \text{Gamma}\left(\frac{\nu+1}{2},\frac{\nu+r_t(\boldsymbol{\beta};\boldsymbol{a})^2/\delta^2}{2}\right). \tag{B1}$$

It follows immediately from (B1) that, if $\epsilon_t \sim$ Cauchy, the weights σ_t are exponentially distributed. For the Modulated Normal type II family, for which $\sigma|\boldsymbol{\zeta} \sim \text{Beta}(\nu/2,1)$, with $\nu > 0$, we have

$$p(\sigma_t|y_t,\boldsymbol{u}_{\boldsymbol{\beta}};\boldsymbol{a},\boldsymbol{\beta},\delta^2,\tau^2) \propto \sigma_t^{(\nu+1)/2-1}\exp\left\{-\frac{r_t(\boldsymbol{\beta};\boldsymbol{a})^2}{2\delta^2}\sigma_t\right\}\mathbb{I}_{(0,1)}(\sigma_t).$$

Although this is similar to the gamma p.d.f., the values of σ_t are constrained to be in the interval $(0,1)$. Hence the simulation is not so straight and we need to use some algorithm like the accept-reject one. However, if $\nu = 1$, we get the Standard Slash distribution (whose pdf is given by $p(x) = (\phi(0)-\phi(x))/x^2$, if $x \neq 0$, and $p(x) = \phi(0)/2$, if $x = 0$, where $\phi(x)$ is the just Standard Normal pdf), for which the full posterior conditional is simplified to $\sigma_t|y_t,\boldsymbol{u}_{\boldsymbol{\beta}};\boldsymbol{a},\boldsymbol{\beta},\delta^2,\tau^2 \sim \text{TEXP}\left(r_t(\boldsymbol{\beta};\boldsymbol{a})^2/(2\delta^2),1\right)$, *i.e.* the Truncated Exponential distribution[1] on the interval $(0,1)$. In the case of the Generalized Hyperbolic distribution, for which the mixing density is such that $p(\sigma|\boldsymbol{\zeta}) \propto \sigma^{-2}\exp\{-0.5(\nu_1/\sigma+\nu_2\sigma)\}$ (here $\boldsymbol{\zeta}' = (\nu_1,\nu_2)$, with $\nu_1 > 0$ and $\nu_2 \geq 0$), we have

$$p(\sigma_t|y_t,\boldsymbol{u}_{\boldsymbol{\beta}};\boldsymbol{a},\boldsymbol{\beta},\delta^2,\tau^2,\boldsymbol{\zeta}) \propto \sigma_t^{(-1/2-1)}\exp\left\{-\frac{1}{2}\left[\left(\frac{r_t(\boldsymbol{\beta};\boldsymbol{a})^2}{\delta^2}+\nu_2\right)\sigma_t+\frac{\nu_1}{2\sigma}\right]\right\},$$

so that

$$\sigma_t|y_t,\boldsymbol{u}_{\boldsymbol{\beta}};\boldsymbol{a},\boldsymbol{\beta},\delta^2,\tau^2,\boldsymbol{\zeta} \sim \text{GIG}\left(\frac{r_t(\boldsymbol{\beta};\boldsymbol{a})^2}{\delta^2}+\nu_2,\nu_1,-\frac{1}{2}\right),$$

[1] The Truncated Exponential distribution with parameter λ on the interval $(0,a)$ is characterized by the pdf $p(x|\lambda) = \lambda\exp\{-\lambda x\}/(1-\exp\{-\lambda a\})$.

where GIG stands for Generalized Inverse Gaussian distribution[2]. In particular, if $\nu_1 = 1$ and $\nu_2 = 0$, we get the Laplace distribution with scale parameter 1. For a more complete list of mixing distributions, we refer the reader to Fernandez *et al.* (2000)

- **Linear Component Parameters:** The posterior full conditional distribution for the linear component parameter $\boldsymbol{\beta}$,

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{u}_{\boldsymbol{\beta}}, \boldsymbol{\sigma}; \boldsymbol{a}, \delta^2, \tau^2) \propto \exp\left\{-\frac{1}{2\delta^2}\mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a})' W^{-1}\mathbf{r}(\boldsymbol{\beta}, \boldsymbol{a})\right\} \mathbb{I}(\boldsymbol{\beta}'\boldsymbol{\beta} = 1), \qquad \text{(B2)}$$

does not match up with any standard distribution, so we sample from (B2) via a Metropolis-Hastings (MH) step by using a von Mises-Fisher distribution, with mean direction $\boldsymbol{\beta}_0$ and concentration parameter $\kappa > 0$, as proposal. Hence, given $\boldsymbol{\beta}_0$,

$$p(\boldsymbol{\beta}_*|\boldsymbol{\beta}_0, \kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)} \exp\{\kappa \boldsymbol{\beta}_0' \boldsymbol{\beta}_*\},$$

where $I_{p/2-1}$ is the Bessel function of first kind[1], and so we take $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, with probability $1 - \rho(\boldsymbol{\beta}_0, \boldsymbol{\beta}^*)$, or $\boldsymbol{\beta} = \boldsymbol{\beta}_*$, with probability $\rho(\boldsymbol{\beta}_0, \boldsymbol{\beta}^*)$, where

$$\begin{aligned}\rho(\boldsymbol{\beta}_0, \boldsymbol{\beta}_*) &= \min\left\{\frac{p(\boldsymbol{\beta}_*|\boldsymbol{y}, \boldsymbol{u}_{\boldsymbol{\beta}}, \boldsymbol{\sigma}; \boldsymbol{a}, \delta^2, \tau^2) p(\boldsymbol{\beta}_0|\boldsymbol{\beta}_*, \kappa)}{p(\boldsymbol{\beta}_0|\boldsymbol{y}, \boldsymbol{u}_{\boldsymbol{\beta}}, \boldsymbol{\sigma}; \boldsymbol{a}, \delta^2, \tau^2) p(\boldsymbol{\beta}_*|\boldsymbol{\beta}_0, \kappa)}, 1\right\} \\ &= \min\left\{\exp\left\{-\frac{1}{2\delta^2}(\mathbf{r}_* - \mathbf{r}_0)' W^{-1}(\mathbf{r}_* + \mathbf{r}_0)\right\} \exp\{\kappa \boldsymbol{\beta}_{prior}'(\boldsymbol{\beta}_* - \boldsymbol{\beta}_0)\}, 1\right\},\end{aligned}$$

is the acceptance probability, $\mathbf{r}_* \equiv \mathbf{r}(\boldsymbol{\beta}_*, \boldsymbol{a})$ and $\mathbf{r}_0 \equiv \mathbf{r}(\boldsymbol{\beta}_0, \boldsymbol{a})$. We have used the identity $p(\boldsymbol{\beta}_*|\boldsymbol{\beta}_0, \kappa) = p(\boldsymbol{\beta}_0|\boldsymbol{\beta}_*, \kappa)$.

Appendix C. Student's t Degrees of Freedom

In order to get the full posterior conditional distribution associated to the exponential prior for ν ($\boldsymbol{\zeta} = \nu$), recall that $\nu \geq \nu_0$ and $\nu - \nu_0 \sim$ Exponencial(λ), with $\lambda \sim$ Uniform(λ_ℓ, λ_u). Hence,

$$\begin{aligned} p(\nu|\boldsymbol{y}, \boldsymbol{u}_{\boldsymbol{\beta}}, \boldsymbol{\sigma}; \boldsymbol{a}, \boldsymbol{\beta}, \delta^2, \tau^2, \lambda) &\propto p(\boldsymbol{\sigma}|\nu) p(\nu|\lambda, \nu_0) \\ &= \left[\prod_{t=1}^{T} \frac{\sigma_t^{\frac{\nu}{2}-1} \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} \exp\left\{-\frac{\nu}{2}\sigma_t\right\}\right] \exp\{-\lambda(\nu - \nu_0)\} \\ &\propto \frac{\left(\frac{\nu}{2}\right)^{\frac{T\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)^T} \exp\left\{-\nu\left(\frac{1}{2}\sum_{t=1}^{T}(\sigma_t - \log\sigma_t) + \lambda\right)\right\} \mathbb{I}(\nu > \nu_0), \end{aligned}$$

and $p(\lambda|\boldsymbol{y}, \boldsymbol{u}_{\boldsymbol{\beta}}, \boldsymbol{\sigma}; \boldsymbol{a}, \boldsymbol{\beta}, \delta^2, \tau^2, \nu) \propto p(\nu|\lambda, \nu_0)\mathbb{I}(\lambda_\ell < \lambda < \lambda_u) = \lambda \exp\{-\lambda(\nu - \nu_0)\} \mathbb{I}(\lambda_\ell < \lambda < \lambda_u)$, which is just a Truncated Gamma distribution. On the other hand, we suggest to sample ν by introducing a Metropolis-Hastings step and for such we take $\nu_* > \nu_0$ such that $\nu_* - \nu_0 \sim LN(\log(\nu_{pr} - \nu_0), \delta_\nu^2)$, *i.e.* the Lognormal distribution , as a proposal (see

[2] The Generalized Inverse Gaussian distribution, GIG(a, b, p), with parameters $a > 0$, $b > 0$ and $p \in (-\infty, \infty)$ on the interval $(0, \infty)$ is characterized by the pdf $p(x|a, b, p) \propto x^{p-1} \exp\{-(ax + b/x)/2\}$.

[1] The Bessel function of first kind and order α is given by $I_\alpha(\kappa) = \sum_{j=0}^{\infty} [j!\Gamma(j + \alpha + 1)]^{-1} (\kappa/2)^{2j+\alpha}$.

Cabral *et al.* (2012)), where ν_{pr} is just the last sampled value of ν. Therefore, using the fact that

$$\frac{p(\nu_{pr}|\log\nu_*,\delta^2_\nu)}{p(\nu_*|\log\nu_{pr},\delta^2_\nu)} = \frac{\nu_*-\nu_0}{\nu_{pr}-\nu_0},$$

we get the acceptance probability

$$\rho(\nu_{pr},\nu_*) = \min\left\{\left[\frac{\Gamma\left(\frac{\nu_{pr}}{2}\right)\left(\frac{\nu_*}{2}\right)^{\frac{\nu_*}{2}}}{\Gamma\left(\frac{\nu_*}{2}\right)\left(\frac{\nu_{pr}}{2}\right)^{\frac{\nu_{pr}}{2}}}\right]^T \times \exp\left\{-\left(\frac{1}{2}\sum_{t=1}^{T}(\sigma_t-\log\sigma_t)+\lambda\right)(\nu_*-\nu_{pr})\right\}\times\right.$$
$$\left.\times\frac{(\nu_*-\nu_0)}{(\nu_{pr}-\nu_0)},1\right\}.$$

References

Andrews, D.R., Mallows C.L., (1974). Scale mixtures of normal distributions. Journal of the Royal Statistical Society, Series B 36, 99-102.

Antoniadis, A., Grégoire G., McKeague, W., (2004). Bayesian estimation in single-index models. Statistica Sinica 14, 1147-1164.

de Boor, C., (2001). A Practical Guide to Splines (2nd Ed.). Springer-Verlag, New York.

Bandyopadhyay, D., Castro, L.M., Lachos, V.H., Pinheiro, H.P., (2015). Robust joint nonlinear mixed-effects models and diagnostics for censored HIV viral loads with CD4 measurement error. Journal of Agricultural, Biological, and Environmental Statistics 20, 121-139.

Cabral, C. R. B., Lachos, V.H., Madruga, M.R., (2012). Bayesian analysis of skew-normal independent linear mixed models with heterogeneity in the random-effects population. Journal of Statistical Planning and Inference 142, 181-200.

Castro, L.M., Lachos, V.H., Ferreira, G.P., Arellano-Valle, R.B., (2014). Partially linear censored regression models using heavy-tailed distributions: A Bayesian approach. Statistical Methodology 18, 14-31.

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

Fang, K., Kotz, S., Ng, K.W., (1989). Symmetric Multivariate and Related Distributions. Chapman and Hall CRC, London.

Fernandez, C., Steel, M.F., (2000). Bayesian regression analysis with scale mixtures of normal. Econometric Theory 16, 80-101.

Fonseca, T.C.O., Ferreira, M.A.R., Migon, H.S., (2008). Objective bayesian analysis for the Student-t regression model. Biometrika 95, 325-333.

Geisser, S., Eddy, W.F., (1979). A predictive approach to model selection. Journal of the American Statistical Association 74, 153-160.

Geweke, J., (1993). Bayesian treatment of the Student's t linear model. Journal of Applied Econometrics 8, S19-S40.

Hastie, T., Tibshirani, R., Friedman, J.H. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York.

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

Müller, P., (1993). Alternatives to the Gibbs sampling scheme. Technical Report, Institute of Statistics and Decision Sciences, Duke University.

Park, C.G., Vannucci, M., Hart, J.D., (2005). Bayesian methods for wavelet series in single-index models. Journal of Computational and Graphical Statistics 14, 1-25.

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

Sheipl, F., Kneib, T., (2009). Locally adaptive Bayesian P-splines with a Normal-Exponential-Gamma prior. Computational Statistics and Data Analysis 53, 3533-3552.

Vehtari, A., Gelman, A., Gabry, J., (2017). Practical Bayesian model evaluation using leave-one-out cross validation and WAIC. Statistics and Computing 27, 1413-1432.

Wang, L., Yang, L., (2009). Spline estimation of single-index models. Statistica Sinica 19, 765-783.

Yu, Y., Ruppert, D., (2002). Penalized spline estimation for partially linear single-index models. Journal of American Statistical Association 97, 1042-1054.