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

Morettin, Pedro
University of São Paulo, Department of Statistics
São Paulo 05508-090, Brazil
pam@ime.usp.br

Taddeo, Marcelo
University of São Paulo, Department of Statistics
São Paulo 05508-090, Brazil
marcelo.taddeo@gmail.com

Introduction

In this paper we consider the single-index model given by

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

for $t = 1, ..., T$, 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 $\mathbf{x}_t$ is a p-dimensional input vector by using bayesian P-splines. We note here that the parameter $\boldsymbol{\beta}$ to be identifiable up to a multiplicative constant, it must have unit (Euclidean) norm, *i.e.*, $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$.

Scale Mixture of Normals

A random variable X is said to follow a scale mixture of Gaussian distributions if it can be written as $X = Z/\sqrt{\sigma}$, where $Z \sim \mathcal{N}(0, 1)$ and σ is any positive (continuous or discrete) random variable. The distribution of σ, H, is called mixture distribution and, if it is absolutelly continuous, its probability density function, h, is called a mixture density. The pdf of scale mixture of normals may be represented by

$$p(x) = \int_0^\infty \sigma^{1/2}\phi(\sigma^{1/2}x)dH_{\boldsymbol{\zeta}}(\sigma), \tag{2}$$

or, when the mixture distribution is absolutelly continuous,

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

where ϕ is the pdf of the standard normal distribution and $\boldsymbol{\zeta}$ stands for the parameter vector associated to the mixture distribution. For a more detailed treatment on the scale mixture of normals, see Andrews and Mallows (1974).

Some distributions which may be modeled as scale mixture of normals and their mixing distributions are: (i) The *Student t distribution* with ν degrees of freedom and scale parameter δ is a scale mixture of normals with $\sigma \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$, i.e., a gamma distribution with both shape and inverse scale parameters (or rate) equal to $\nu/2$. (ii) The *Cauchy distribution* is the particular case where $\nu = 1$. (iii) This class also contains the *hyperbolic* and *variance-gamma distributions*, the logistic and contaminated normal distributions and the modulated normal distribution type II (the Slash distribution being a particular case).

From a modeling perspective, we note that if we write

$$\begin{aligned} y_t|\mathbf{x}_t; \sigma_t &\sim \mathcal{N}\left(f(\mathbf{x}_t), \frac{\delta^2}{\sigma_t}\right), \\ \sigma_t &\sim h_{\boldsymbol{\zeta}}, \end{aligned} \tag{3}$$

where δ stands for a scale parameter, then the pdf of y_t is given by

$$(4) \quad p(y_t|\mathbf{x}_t) = \int_0^\infty \frac{\sigma_t^{1/2}}{\delta}\phi\left(\frac{\sigma_t^{1/2}}{\delta}(y_t - f(\mathbf{x}_t))\right) h_{\boldsymbol{\zeta}}(\sigma_t)d\sigma_t,$$

which is just a scaled and shifted version of (2). The random variables σ_t are not observable and so they must be treated as latent variables.

Splines, B-Splines and P-Splines

We assume that the (link) function g can be writtes as a linear combination of some basis functions such as the B-splines,

$$g(x) = \sum_{i=1}^{M} a_i B_i(x),$$

where M is the number of elements in the basis and B_i is the its i*th* member. Thus, $(g(\boldsymbol{\beta}'\mathbf{x}_1), ..., g(\boldsymbol{\beta}'\mathbf{x}_T))' = B\mathbf{a}$, with $B(\boldsymbol{\beta})$ being a matrix with $(B_i(\boldsymbol{\beta}'\mathbf{x}_1), \ldots, B_i(\boldsymbol{\beta}'\mathbf{x}_T))'$ as its i-th column and therefore

$$\mathbf{y} = B(\boldsymbol{\beta})\mathbf{a} + \delta\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_T)'$. Notice that, conditioning $\boldsymbol{\epsilon}$ on the latent variables σ_t, we get

$$(5) \quad \mathbf{y}|\boldsymbol{\sigma}; \mathbf{a}, \boldsymbol{\beta}, \delta^2 \sim \mathcal{N}\left(B(\boldsymbol{\beta})\mathbf{a}, \delta^2 W\right),$$

where $\sigma_t \overset{\text{iid}}{\sim} p_\sigma(\sigma_t|\boldsymbol{\zeta})$, $W = \text{diag}(\sigma_1^{-1}, ..., \sigma_t^{-1})$. Here, $\boldsymbol{\zeta}$ is just a parameter vector which determines the prior distribution of $\boldsymbol{\sigma}$.

We impose some penalizations on the estimate of the target function. Here, such penalizations are introduced by using *P-splines*, which are just the combination of B-splines and the penalization on $\mathbf{a}$ given by

$$(6) \quad \lambda \sum_{j=d+1}^{M} (\Delta^d a_j)^2,$$

where λ acts as a smoothing parameter. In fact, for large values of λ, the target function estimate is smoother. This approach is computationally attractive and can be easily written in the matrix form of the operator difference of order d, which we shall represent by K_d. For more on the construction of the matrix K_d, we refer to Eilers and Marx (1996) and Eilers and Marx (2010).

Bayesian P-Splines

In the Bayesian approach the penalties in (6) are replaced by their stochastic counterparts. For example, when the first differences Δa_j are penalized, we consider a first order random walk. On the other hand, if we consider second differences, a second-order random is used. More precisely, we assume

$$a_j = a_{j-1} + u_j$$

and

$$a_j = 2a_{j-1} - a_{j-2} + u_j,$$

where $\{u_j\}$ is a white noise, respectively. As in Lang and Brezger (2004), we 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. We note that τ^2 works as a smoothing parameter and has a role similar to the smoothing parameter λ used in the frequentist case. A consequence is that

$$(7) \quad p(\mathbf{a}|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\mathbf{a}'K\mathbf{a}\right),$$

where K is equal to K_1 and K_2 for the first-order and second-order random walk, respectively. It is worth to note that conditioning $\mathbf{y}$ on the latent random vector $\boldsymbol{\sigma}$, the relation (5) holds and

$$\begin{aligned}\log p(\mathbf{a}, \boldsymbol{\beta}, \delta^2, \tau^2; \boldsymbol{\sigma}, \boldsymbol{\zeta}|\mathbf{y}) \cong \frac{1}{2\delta^2}(\mathbf{y} - B(\boldsymbol{\beta})\mathbf{a})'W^{-1}(\mathbf{y} - B(\boldsymbol{\beta})\mathbf{a}) - \frac{1}{2\tau^2}\mathbf{a}'K\mathbf{a} \\ + \log p_\beta(\boldsymbol{\beta}) + \log p_\delta(\delta^2) - \log \delta^2 + \log p_\tau(\tau^2) \\ + \log h(\boldsymbol{\sigma}|\boldsymbol{\zeta}) + \frac{1}{2}\sum_{t=1}^{T} \log \sigma_t + \log p_\zeta(\boldsymbol{\zeta}),\end{aligned}$$

where $\cong$ stands for equality up to a constant. Hence the posterior maximum likelihood estimate of $\mathbf{a}$ corresponds to a penalized weighted least-squares estimate.

Likelihood and Priors

We have already set priors to the B-splines coefficients $\mathbf{a}$ so that

$$\mathbf{a}|\tau^2 \sim \mathcal{N}(\mathbf{0}, \tau^2 K_d^{-1}).$$

In what follows we set the other priors.

Likelihood and Modeling by Using Latent Variables

Modeling the data in $\mathbf{y}$ as in (3), we get

(8) $$\mathbf{y}|\boldsymbol{\sigma} \sim \mathcal{N}\left(B_{\boldsymbol{\tau}}(\mathbf{x}; \boldsymbol{\beta})\mathbf{a}; \delta^2 W\right),$$

where $\sigma_t \overset{\text{iid}}{\sim} h$ and $W = \left(\frac{1}{\sigma_1}, ..., \frac{1}{\sigma_T}\right)$. Of course, the above r.v.s are conditioned on the input variables $\mathbf{x}_t$. As we shall see the r.v.s σ_t work as weights which neutralize the effects of an eventual heavy tail phenomenon. Besides, such hierarchical structure, i.e., using the latent variables σ_t, turns the data analysis much simpler.

Priors

Index Vector (β): we take as prior distribution the uniform distribution (on the hypersphere embedded in $\mathbb{R}^p$, $\{\boldsymbol{\beta} = (\beta_1, ..., \beta_p) : \|\boldsymbol{\beta}\| = 1\}$, see Wang (2009)) so that

$$p(\boldsymbol{\beta}) = \pi^{-p/2}\Gamma\left(\frac{p}{2}\right).$$

Scale and Smoothing Parameters (δ^2 and τ^2): We assume a gamma-inverse prior distribution for the scale and smoothing parameters. More precisely, we set $\delta^2 \sim \text{GI}(\alpha_0, \gamma_0)$ and $\tau^2 \sim \text{GI}(\alpha_1, \gamma_1)$.

Mixture Distribution Parameters (ζ): In this paper, we give special attention to Student-t distribution and set an exponential prior for $\zeta = \nu$,

$$p(\nu) = \lambda \exp\{-\lambda\nu\}.$$

Sampling Scheme

In order to sample the parameters from their joint posterior distribution, we suggest to use the Metropolis-within-Gibbs algorithm. In fact, it is not possible to represent all the full conditional distributions for the parameters $\mathbf{a}$, $\boldsymbol{\beta}$, δ^2, τ^2 and ν as well as the full conditional distribution for the latent variables σ_t in terms of standard distributions.We shall consider $\boldsymbol{\zeta}$ as known, but later we shall consider the particular case where the degrees of freedom ($\zeta = \nu$) for the Student's t distribution is not known.

Some of the full conditional distributions may be fully determined, namely $p(\mathbf{a}|\mathbf{y}, \mathbf{x}, \boldsymbol{\sigma}; \boldsymbol{\beta}, \delta^2, \tau^2)$, $p(\delta^2|\mathbf{y}, \mathbf{x}, \boldsymbol{\sigma}; \mathbf{a}, \beta$ and $p(\tau^2|\mathbf{y}, \mathbf{x}, \boldsymbol{\sigma}; \mathbf{a}, \boldsymbol{\beta}, \delta^2)$. See Taddeo and Morettin for details.

Metropolis-within-Gibbs

Until now we have been able to explicitly derive the full conditional distributions associated with some of the parameters, but the same is not necessarily true for $\boldsymbol{\sigma}$, $\boldsymbol{\beta}$ and, when appropriate, $\boldsymbol{\zeta}$. To overcome this difficulty, we introduced a Metropolis-Hastings step into the Gibbs sampler. For this, we could, for example, use a prior proposal ($\sigma_t^* \sim h(\sigma_t|\boldsymbol{\zeta})$). However, although this procedure is quite simple and general, if the likelihood p and the prior (and proposal) h are not concentrated in the same region, this method may be not very efficient, and therefore a more sophisticated approach would be necessary. On the other hand, fortunately, there are some interesting cases where the above posterior distribution may written as a closed formula: the contaminated normal, the Student's t and the Modulated Normal type II family distributions. In the first case, we have

$$p(\sigma_t|y_t, \mathbf{x}; \mathbf{a}, \boldsymbol{\beta}, \delta^2, \tau^2) = \begin{cases} 1 - \xi', & \text{if } \sigma = 1, \\ \xi', & \text{if } \sigma = \lambda^2, \\ 0, & \text{otherwise,} \end{cases}$$

where $\xi' = C_t \xi \lambda \exp\{-r_t(\boldsymbol{\beta}; \mathbf{a})^2 \lambda/(2\delta^2)\}$ and C_t is a normalizing constant which depends on the observation $(y_t; \mathbf{x}_t')'$. For the Student's t distribution, we have

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

so that

$$\text{(9)} \quad \sigma_t|y_t, \mathbf{x}; \mathbf{a}, \boldsymbol{\beta}, \delta^2, \tau^2 \sim \Gamma\left(\frac{\nu+1}{2}, \frac{\nu + r_t(\boldsymbol{\beta}; \mathbf{a})^2/\delta^2}{2}\right).$$

It follows immediately from (9) that, if $\epsilon_t \sim$ Cauchy, the weights σ_t are exponentially distributed. For the Modulated Normal type II family, we have

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

Although this is similar to the p.d.f. associated to the gamma distribution, the values of σ_t are constrained to be in the interval $(0, 1)$. Hence in this case the simulation is not straight and we need to use some algorithm like, for example, the accept-reject one. Finally, for the hyperbolic and for the variance-gamma distributions,

$$\sigma_t|y_t, \mathbf{x}; \mathbf{a}, \boldsymbol{\beta}, \delta^2, \tau^2 \sim \mathcal{GIG}\left(\psi + \frac{r_t(\boldsymbol{\beta}; \mathbf{a})^2}{\delta^2}, \kappa, -1\right),$$

and

$$\sigma_t|y_t, \mathbf{x}; \mathbf{a}, \boldsymbol{\beta}, \delta^2, \tau^2 \sim \mathcal{GIG}\left(\frac{r_t(\boldsymbol{\beta}; \mathbf{a})^2}{\delta^2}, \nu, \frac{1-\nu}{2}\right),$$

respectively.

In the case of the parameter $\boldsymbol{\beta}$, note that the (posterior) full conditional distribution is given by

$$\text{(10)} \quad p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \boldsymbol{\sigma}; \mathbf{a}, \delta^2, \tau^2) \propto \exp\left\{-\frac{1}{2\delta^2}\mathbf{r}(\boldsymbol{\beta}, \mathbf{a})' W^{-1} \mathbf{r}(\boldsymbol{\beta}, \mathbf{a})\right\} \mathbb{I}(\boldsymbol{\beta}'\boldsymbol{\beta} = 1).$$

Since this distribution does not match up with any standard distribution, we use a Metropolis-Hastings step to sample from (10). To this end, we assume a von Mises-Fisher distribution, with mean direction $\boldsymbol{\beta}_0$ and concentration parameter $\kappa > 0$, as proposal, so that, given $\boldsymbol{\beta}_0$ (sampled in the previous step of the algorithm),

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

where $I_{p/2-1}$ denotes the Bessel function of first kind and order $p/2-1$. Finally, we take

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

where

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

is the acceptance probability and $\mathbf{r}_* \equiv \mathbf{r}(\boldsymbol{\beta}_*, \mathbf{a})$ and $\mathbf{r}_0 \equiv \mathbf{r}(\boldsymbol{\beta}_0, \mathbf{a})$. In the above result, we have used the fact that $p(\boldsymbol{\beta}_*|\boldsymbol{\beta}_0, \kappa) = p(\boldsymbol{\beta}_0|\boldsymbol{\beta}_*, \kappa)$.

Sampling the Degrees of Freedom for Student t Errors

Given that we are paying particular attention to the Student-t distribution, we briefly describe how $\boldsymbol{\zeta} = \nu$ may be sampled from its posterior distribution (for details, we refer to Geweke (1993) and the references therein). The target distribution is

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

and the proposal distribution is given by an exponential distribution with parameter λ_*, which is a parameter to be calibrated. Note that in this case, we would have to use twice the Metropolis-Hastings algorithm. However, as suggested by Müller, see Müller (1993), we can reduce the acceptance or rejection of the proposals in a single step. This results in one Metropolis-Hastings algorithm within the Gibbs sampling, rather than a combination of such algorithms and also produces a global approximation (for the posterior full conditional of $\boldsymbol{\beta}$ and ν), instead of local approximations of the individual posterior full conditional distributions of $\boldsymbol{\beta}$ and ν, respectively. Now, since the proposals for $\boldsymbol{\beta}_*$ and ν_* are independent and

$$p(\boldsymbol{\beta}, \nu|\mathbf{y}, \mathbf{x}, \boldsymbol{\sigma}; \mathbf{a}, \delta^2, \tau^2) = p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \boldsymbol{\sigma}; \mathbf{a}, \delta^2) p(\nu|\boldsymbol{\sigma})$$

it follows that the "joint" acceptance probability is given by

$$\begin{aligned} \rho((\boldsymbol{\beta}_0', \nu_{pr})', (\boldsymbol{\beta}_*', \nu_*)') &= \min\Bigg\{ \exp\left\{ -\frac{1}{2\delta^2} (\mathbf{r}_* - \mathbf{r}_0)' W^{-1} (\mathbf{r}_* + \mathbf{r}_0) \right\} \\ &\qquad \exp\Bigg\{ -(s_t - \lambda_*)(\nu_* - \nu_{pr}) \Bigg\} \left[\frac{\left(\frac{\nu_*}{2}\right)^{\frac{\nu_*}{2}} \Gamma\left(\frac{\nu_{pr}}{2}\right)}{\left(\frac{\nu_{pr}}{2}\right)^{\frac{\nu_{pr}}{2}} \Gamma\left(\frac{\nu_*}{2}\right)} \right]^T \\ &= \min\{\gamma(\boldsymbol{\beta}_0, \boldsymbol{\beta}_*) \gamma(\nu_{pr}, \nu*), 1\}, \end{aligned}$$

where

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

and

$$\gamma(\nu_{pr}, \nu*) \equiv \exp\left\{(s_t\nu_{pr} - 1)\left(1 - \frac{\nu_*}{\nu_{pr}}\right)\right\}\left[\frac{\left(\frac{\nu_*}{2}\right)^{\frac{\nu_*}{2}}\Gamma\left(\frac{\nu_{pr}}{2}\right)}{\left(\frac{\nu_{pr}}{2}\right)^{\frac{\nu_{pr}}{2}}\Gamma\left(\frac{\nu_*}{2}\right)}\right]^T,$$

with

$$s_t \equiv \frac{1}{2}\sum_{t=1}^{T}(\sigma_t - \log\sigma_t) + \lambda.$$

Remarks

See Taddeo and Morettin (2011) for a simulation study and an application to an environmental study. Single-Index models are a way to overcome the well-known Curse of Dimensionality, typical of nonparametric multivariate models. In this paper, we chose to allow the noise to follow distributions belonging to a broader class of distributions than just members of the class of normal distributions, the class of scale mixture of Gaussians distributions. We present a Bayesian framework for estimation and inference of the parameters of interest, *i.e.*, the parameters that make up the model (1).

References

D. R. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *J. R. Statist. Soc. B*, 36:99–102, 1974.

Paul Eilers and Brian D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2): 89–121, 1996.

Paul Eilers and Brian D. Marx. Splines, knots and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2010.

John Geweke. Bayesian treatment of the student's-t linear model. *Journal of Applied Econometrics*, 8:S19–S40, 1993.

Stefan Lang and Andreas Brezger. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13: 183–212, 2004.

Peter Müller. Alternatives to the Gibbs sampling scheme. Technical report, Institute of Statistics and Decision Sciences, Duke University, 1993.

M. Taddeo and P.A. Morettin. The analysis of single-index models with scale mixture of normals errors by using bayesian p-splines. *TR, IME-USP*, 2011.

Hai-Bin Wang. Bayesian estimation and variable selection for single index models. *Journal of Computational and Graphical Statistics*, 53:2617–2627, 2009.