Estimation of Semiparametric Models with Errors Following a Scale Mixture of Gaussian Distributions

Marcelo M. Taddeo\textsuperscript{1} and Pedro A. Morettin\textsuperscript{2}

\textsuperscript{1} Federal University of Bahia, Brazil and \textsuperscript{2} University of São Paulo, Brazil

Abstract

In this paper we consider a semiparametric regression model where the error follows a scale mixture of Gaussian distributions. The purpose is to estimate the target function which is assumed to belong to some class of functions using the EM algorithm and approximations via \( P \)-splines and \( B \)-splines. We illustrate the proposed methodology through several simulation studies. Other forms of function approximation are also studied, namely Fourier and wavelet expansions.

Key words: semiparametric model, regression, EM algorithm, \( P \)-splines, \( B \)-splines, wavelets.

1 Introduction

In this paper we consider models of the type

\[ y_t = f(x_t) + \delta \epsilon_t, \]

for \( t = 1, ..., T \), where \( x_t \) can be deterministic, random and, in principle, equal to a lagged realization of a stochastic process \( Y \), e.g. \( y_{t-1} \). Besides that, \( \delta \) is a scale parameter and the error \( \epsilon_t \) is assumed to follow a scale mixture of Gaussian distributions, i.e.

\[ p(\epsilon_t) = \int_0^\infty \frac{1}{\psi(s)} \phi \left( \frac{\epsilon_t}{\psi(s)} \right) dH_\zeta(s). \]

Here \( H \) is a distribution over \([0, \infty)\) possibly depending on a parameter vector \( \zeta \), \( \phi \) is the standard Normal density and \( \psi \) is a function known \textit{a priori}, see Andrews and Mallows (1974) and West (1987). We indicate by \( \epsilon_t \sim SM_h(\delta, \psi) \) if the density is given by (2).

Our goal is to estimate the target function \( f \), which is assumed to belong to some class of functions \( \mathcal{H} \). Traditionally, model (1) has been estimated using kernels, splines and more recently wavelets. The case of equally spaced observations was considered by Donoho and Johnstone (1994) with \( \epsilon_t \) Gaussian. This was further extended to the case of stationary Gaussian errors by Johnstone and Silverman (1997). The case of
irregular design and Gaussian errors was treated by Cai and Brown (1998) and for correlated errors by Porto et al. (2008). In all these cases we had fixed designs. For $x_t$ uniform in $[0, 1]$ and Gaussian errors, see Cai and Brown (1999). Porto et al. (2014) extended this situation for the case of correlated errors. In this article, we study the estimation of semiparametric models with additive error distributed according to a mixture of normals. Some classes of models with different penalties are considered. The evaluation of estimators and ways of dealing with complicated distributions via Monte-Carlo are also discussed. The methods are evaluated and compared via simulation studies. An application to real data is performed to illustrate the method.

Scale mixtures of Gaussians are interesting since they include a huge number of continuous symmetric densities, see Fang (1989), p.50. In particular, it includes many heavy tailed distributions like the Student’s $t$, the Double Exponential and the symmetric stable distributions. From (1) and (2), we have

$$p(y_t|x_t) = \int_0^\infty \frac{1}{\delta \psi(s)} \phi \left( \frac{y_t - f(x_t)}{\delta \psi(s)} \right) dH_\zeta(s).$$

If $\sigma_t \sim H_\zeta$, then $\epsilon_t|\sigma_t \sim \mathcal{N}(0, \psi(\sigma_t)^2)$. To keep notation simpler, from now on we will write $p(y_t)$ instead of $p(y_t|x_t)$, the dependency on $x_t$ becoming clear by the context. The vector of parameters, denoted by $\theta$, is assumed to belong to $H' \times B \subseteq H \times \mathbb{R}^d$, where $H'$ is a subspace of $H$.

If the random variable $\sigma$ were observable, then we could estimate the target function $f$ by maximizing

$$-\frac{1}{T} \sum_{t=1}^T \left( \frac{y_t - f(x_t)}{\delta^2 \psi(\sigma_t)} \right)^2 + J_\lambda(f),$$

where $J_\lambda$ is a penalizing function and $\lambda$ a penalization parameter, or through a penalized maximum likelihood method weighted by the vector $(\psi(\sigma_1), ..., \psi(\sigma_T))$. This vector weights atypical observations, guaranteeing the robustness of the model. If we consider, for instance, the penalty $J_\lambda(f) = -\lambda \int [f''(x)]^2 dx$ and allow $f$ to be in a Sobolev space $W_2^m$, then the maximizer of (3) is a spline (with knots in every observation). This is the idea supporting the smoothing splines approach. The larger the overall curvature of the curve $f$, the larger is the penalization $J_\lambda$, so that it is a measure of the roughness associated to $f$. Indeed, $J_\lambda$ favors smoother functions and the parameter $\lambda$ governs the tradeoff between smoothness and bias of the estimator of $f$. If $\lambda = 0$, the estimation procedure results in data interpolation, while for $\lambda \to \infty$ we get a linear estimate. It is worthwhile to point out that, defining $H$ as a Sobolev space, the optimization problem (3) which is infinite dimensional, reduces to a problem of finite dimension, given that the optimal solution of the problem is a cubic spline. On the other hand, if $J_\lambda \equiv 0$, then we are back to the usual least squares or maximum likelihood method. Another way to estimate $f$ and reduce the problem to a finite (and small) dimension is to approximate it by a linear combination of basis functions, say $B_j$ ($j = 1, ..., M$), such as wavelets or B-splines. The problem, then, reduces to finding the vector of coefficients $(c_1, ..., c_M)^\top$ that maximizes the corresponding objective function. It turns out, however, that the calculation of derivatives and integrals in the penalization function $J_\lambda$ is not trivial, especially for orders greater than 2. A very usual alternative is the use of $P$-splines, in which the penalty term is based on differences in a certain order of the coefficient vector associated to the $B$-splines approximation, see Eilers and Marx (1996).
However, in this paper \( \sigma \) arises from the scale mixture of Normals and is a latent variable, so that the likelihood full structure must be taken into account. The estimates of the target function \( f \) and of the scale parameter \( \delta \) should maximize
\[
\frac{1}{T} \sum_{t=1}^{T} \ell(f, \delta, \zeta; y_t) + J(\lambda(f)),
\]
where \( \ell(f, \delta, \zeta; y_t) \) is the log-likelihood associated to \( y_t \). In the following sections, we will discuss how to obtain the estimates of \( f \) and \( \delta \) which maximize (4).

The paper is organized as follows. In Section 2 we describe all applications of the EM algorithm using smoothing splines, B-aplines and P-splines. All algorithms and technical issues, which include bootstrapping, construction of confidence bands, the use of Monte-Carlo method in specific situations and extensions to multivariate cases, are described in this section. Section 3 is dedicated to simulation studies, analysis of real data and some notes on the applicability of the methodology with trigonometric series and wavelets. In Section 4, we conclude the paper.

2 Estimation Algorithm

Assume \( Y_t \sim SM_h(f(x_t), \delta; \psi) \) and that it can be represented in the hierarchical form
\[
y_t|\sigma_t \sim N(f(x_t), \delta^2 \psi(\sigma_t)^2),
\]
\[
\sigma_t \sim h.
\]

To estimate the parameters of interest and to incorporate the penalization function, we propose to use the modified EM algorithm (see Green and Silverman, 1994). The latent variables \( \sigma_t \) control the variance of \( y_t \) and, hence, smooth the impact of outliers on the estimators. By restricting the space where \( f \) lies and finding the parameters that maximize the log-likelihood of the model (5), we also obtain the parameters that maximize the log-likelihood \( \ell \), Casella and Roberts (1994). Before we continue, we set some notation: \( y = (y_1, ..., y_T)^\top \) and \( z = (z_1, ..., z_T)^\top \), with \( z_t = (y_t, x_t)^\top \), are the vectors of observations; and \( \sigma = (\sigma_1, ..., \sigma_T)^\top \) is the vector of latent variables. The main component of the objective function to be defined below is \( Q(\theta|\theta') = \mathbb{E}_{\theta'}\{\log p(y, \sigma|\theta)|z\} \), where (i) the expectation is with respect to the distribution of \( \sigma \) given \( z \) and using the vector of parameters \( \theta' \), and (ii) \( p \) is the joint distribution of \( (y_1, \sigma_1), ..., (y_T, \sigma_T) \) (given \( (x_1, ..., x_T)^\prime \)). Particularly, we see that \( Q(\theta|\theta') = \sum_{t=1}^{T} Q_t(\theta|\theta') \), where \( Q_t(\theta|\theta') = \mathbb{E}_{\theta'}\{\log p(y_t, \sigma_t|\theta)|z_t\} \).

Assume that the elements of \( \mathcal{H}' \) can be written as linear combinations of some basis functions \( \{B_i\}_{i=1}^{K} \), where \( K \in \mathbb{Z} \) or \( K = \infty \). For example, if we assume that \( \mathcal{H}' \) is the space of polynomials of degree less or equal \( p \), then we could take \( K = p \) and \( B_j(x) = x^j \), for \( j = 0, 1, ..., p \). If we assume that \( \mathcal{H}' \) is the Sobolev space \( W_2^2[0, 1] \), we could take \( K = \infty \) and the \( B_j \)'s elements of any complete orthonormal sequence. Anyway,
the algorithm consists in maximizing the penalized log-likelihood by iteratively maximizing

$$S(\theta|\theta') \equiv Q(\theta|\theta') - J_\lambda(\theta) = \sum_{t=1}^{T} Q_t(\theta|\theta') - J_\lambda(\theta)$$

with respect to $\theta = (f, \delta, \zeta)$ until convergence. Note that

$$\log p(y_t, \sigma_t|\theta) = \log \left[ \frac{1}{\delta \psi(\sigma)} \phi \left( \frac{y_t - f(x_t)}{\delta \psi(\sigma)} \right) \right] + \log h_\zeta(\sigma_t)$$

so that

$$Q_t(\theta|\theta') = E_{\theta'} \left\{ \log \left[ \frac{1}{\delta \psi(\sigma)} \phi \left( \frac{y_t - f(x_t)}{\delta \psi(\sigma)} \right) \right] + \log h_\zeta(\sigma_t) \right\}$$

where

$$\log \left[ \frac{1}{\delta \psi(\sigma)} \phi \left( \frac{y_t - f(x_t)}{\delta \psi(\sigma)} \right) \right] = -\frac{1}{2} \log(2\pi) - \log \psi(\sigma_t) - \log \delta - \frac{(y_t - f_\theta(x_t))^2}{2\delta^2 \psi(\sigma_t)^2}.$$ 

Therefore,

$$Q_t(\theta|\theta') = -\log \sqrt{2\pi} - E_{\theta'} \left[ \log \psi(\sigma_t)|z_t \right] - \log \delta - E_{\theta'} \left[ \frac{1}{\psi(\sigma_t)^2} \right] (y_t - f_\theta(x_t))^2 + E_{\theta'} \left[ \log h(\sigma_t|\theta)|z_t \right].$$

(6)

Since our aim is to maximize $Q(\theta'|\theta)$, we can ignore the term $-\log \sqrt{2\pi}$ in (6) and consider

$$Q_t(\theta|\theta') = -E_{\theta'} \left[ \log \psi(\sigma_t)|z_t \right] - \log \delta - E_{\theta'} \left[ \frac{1}{\psi(\sigma_t)^2} \right] (y_t - f_\theta(x_t))^2 + E_{\theta'} \left[ \log h(\sigma_t|\theta)|z_t \right].$$

Using the fact that $y_1, ..., y_T$ are independent,

$$Q(\theta|\theta') = -T \log \delta - \frac{1}{2\delta^2} \sum_{t=1}^{T} E_{\theta'} \left[ \frac{1}{\psi(\sigma_t)^2} \right] (y_t - f_\theta(x_t))^2 + C(\zeta; \theta'),$$

(7)

where

$$C(\zeta; \theta') \equiv \sum_{t=1}^{T} E_{\theta'} \left[ \log \frac{h(\sigma_t|\theta)}{\psi(\sigma_t)} \right] \left| z_t \right|,$$

which contains only the parameters associated to the distribution of the error. Finally, writing

$$W_T(\theta') = \text{diag} \left( E_{\theta'} \left[ \psi(\sigma_1)^{-2}|z_t \right], ..., E_{\theta'} \left[ \psi(\sigma_T)^{-2}|z_t \right] \right),$$

(8)

the weighting vector for the estimation of $f$ and $y - f = (y_1 - f(x_1), ..., y_T - f(x_T))^T$, we get

$$Q(\theta|\theta') = C(\zeta; \theta') - T \log \delta - \frac{1}{2\delta^2} \left( y - f \right)^T W_T(\theta')(y - f).$$

(9)
We notice that (i) the matrix $W_T$ depends on $\theta$ only if $\psi$ does and (ii) $C$ independs of $f$ and of parameters associated to its approximation. In the presence of the penalization term $J_\lambda$, the objective function becomes

$$S(\theta | \theta') = C(\zeta; \theta') - T \log \delta - \frac{1}{2\delta^2}(y - f)^\top W_T(\theta')(y - f) - J_\lambda(f)$$  \hspace{1cm} (10)$$

and, as we mentioned before, the representation of $J_\lambda$ depends on hypotheses made about the space to which $f$ belongs and about the basis used to approximate $f$. These specifications will be dealt with in the next section.

Regarding the expectations above, they are computed relative to the conditional distributions

$$k(\sigma_t | z_t, \theta') = \frac{p(\sigma_t, y_t | \theta')}{p(y_t | \theta')} \propto \frac{1}{\delta \psi(\sigma_t)} \phi \left( \frac{y_t - \hat{f}_\theta(x_t)}{\delta \psi(\sigma_t)} \right) h(\sigma_t | \theta')$$

$$\approx \frac{1}{\delta \psi(\sigma_t)} \phi \left( \frac{y_t - \hat{f}_\theta(x_t)}{\delta \psi(\sigma_t)} \right) h(\sigma_t | \theta'),$$  \hspace{1cm} (11)$$

where $\hat{f}_\theta$ is the approximation of $f$ based in the parameters estimated in the previous iteration. In the instances in which the integrals in (7) are analytically intractable, we some numerical method must be used in the E-step of the EM algorithm.

We also notice that it usual to consider $\psi(\sigma) = \sigma^{-1/2}$. In such a case, the weighting matrix becomes $W_T(\theta') = \text{diag}(E_{\theta'}[\sigma_1 | z_t], ..., E_{\theta'}[\sigma_T | z_t])$, and we may take $C(\zeta; \theta') = \sum_{t=1}^T E_{\theta'} \{ \log h(\sigma_t | \theta | z_t) \}$, since $\psi$ does not depend on $\theta$ anymore. Moreover, (11) becomes

$$k(s | y; \theta) \propto \frac{\sqrt{s}}{\delta} \phi \left( \frac{\sqrt{s}}{\delta} (y - f_\theta(x)) \right) h(s | \theta) \approx \frac{\sqrt{s}}{\delta} \phi \left( \frac{\sqrt{s}}{\delta} (y - \hat{f}_\theta(x)) \right) h(s | \theta).$$  \hspace{1cm} (12)$$

We now consider some special cases.

**Example 1** (Student’s $t$, $\nu$ known). Let $\epsilon_t \sim t_\nu(0, \zeta^2)$ ($\nu$ known), so that $y_t \sim t_\nu(f(x_t), \delta^2)$. If $y_t | \sigma_t \sim \mathcal{N}(f(x_t), \delta^2/\sigma_t)$ and $\sigma_t \sim \text{Gamma}(\nu/2, 2/\nu)$ Since $\nu$ is known and we are considering $\psi(\sigma) = 1/\sqrt{\sigma}$, we have

$$Q(\theta | \theta') = -T \log \delta - \frac{1}{2\delta^2}(y - f)^\top W_T(\theta')(y - f),$$  \hspace{1cm} (13)$$

with $E_{\theta'}[\psi(\sigma_t)^{-2} | y_t] = E_{\theta'}[\sigma_t | y_t] = \tilde{\sigma}_t$ and $W_T(\theta') \equiv \text{diag}\{\tilde{\sigma}_1, ..., \tilde{\sigma}_T\}$. From (11) and the definition of the gamma density, $k(\sigma_t | y_t, \theta')$ is determined by $\sigma_t | y_t, \theta' \sim \text{Gamma}(\nu + 1/2, (\nu + r_t^2/\delta^2)/2)$, where $r_t = y_t - \tilde{f}_t$. In particular, the weights $\tilde{\sigma}_t$ are given by $\tilde{\sigma}_t = (\nu + 1)(\nu + r_t^2/\delta^2)$. Notice that the bigger the distance between an observation and the fitted curve, the smaller will be the weight given to this observation in the estimation of the parameters of the model.

**Example 2** (Cauchy). The Cauchy distribution is a special case of the Student $t$ with $\nu = 1$. Then it follows that $\tilde{\sigma}_t = 2(1 + r_t^2/\delta^2)^{-1}$.  

5
Example 3 (Double Exponential). When the errors $\varepsilon_t$ follow a Double Exponential (Laplace) distribution, so that $p(\varepsilon_t) \propto \exp\{-|\varepsilon_t|\}$, then the mixture density is given by $h(\sigma) = \exp\{-1/(2\sigma^2)\}/(2\sigma^2)$ and the corresponding weights are $\bar{\sigma}_t = \delta/|r_t|$. 

Example 4 (Contaminated normal). Given the mixture density

$$h(\sigma) = \begin{cases} 
\pi_0, & \text{if } \sigma = 1 \\
\pi_j, & \text{if } \sigma = s_j^2 \text{ for } j = 1, \ldots, K \\
0, & \text{otherwise}
\end{cases}$$

where $\pi_0, \ldots, \pi_K > 0$ and $\sum_{j=0}^{K} \pi_j = 1$, we have the model $y_t|\sigma_t \sim \mathcal{N}(f(x_t), \delta^2/\sigma_t)$ and $p(\sigma_t = s_j^2) = \pi_j$, for $j = 0, 1, \ldots, K$, and $s_0 = 1$. Specifically, $\bar{\sigma}_t = \sum_{j=0}^{K} \pi_j s_j$, where $\pi_j s_j = C_t^{-1} k(\sigma_t = s_j^2|y_t, \theta')$, with

$$k(\sigma_t = s_j^2|y_t, \theta') \propto \frac{s_j \phi \left( \frac{r_t}{\delta s_j} \right)}{\pi_j},$$

and $C_t = \delta^{-1} \sum_{j=0}^{K} s_j \phi(s_j r_t/\delta) \pi_j$.

If we approximate $f(x)$ by $\sum_{j=1}^{M} c_j B_j(x)$, with $c = (c_1, \ldots, c_M)^\top$ and the functional $J_\lambda(f)$ by $\tilde{J}_\lambda(c) = \lambda c^\top \Omega c$, where $\Omega$ is some positive-definite matrix, then

$$S(\theta|\theta') = C(\zeta; \theta') - T \log \delta - \frac{1}{2\delta^2} (y - Bc)^\top W_T(\theta')(y - Bc) - \lambda c^\top \Omega c,$$

where $\theta^\top = (c^\top, \delta, \zeta^\top)$ and $B = [B_j(x_i)]$ is the design matrix. In the next sections we discuss the estimation of $c$ under different approximation strategies with their corresponding design matrices and penalization matrix $\Omega$. However, in all cases, the estimators of $\delta$ and $\zeta$ remain the same. Indeed, since $\tilde{J}_\lambda$ does not depend on them, such estimators are obtained by differentiating $Q$ with respect to them and equating the result to zero. 

Regarding the scale parameter, we get

$$\tilde{\delta}^2 = \frac{1}{T} (y - Bc)^\top W_T(\theta')(y - Bc),$$

while for $\zeta$, if $\psi$ does not depend on $\zeta$, its estimator is the solution of $\frac{\partial}{\partial \zeta} S(\theta|\theta') = \frac{\partial}{\partial \zeta} C(\zeta; \theta') = 0$.

Finally, it is interesting to note that, from a computational point of view, there is no impediment to considering time series, allowing the $x_t$ regressor to be the lagged variable $y_{t-k}$ if there is reason to believe that, conditionally on $y_{t-k}$, the variable $y_t$ is conditionally independent of the other elements of the series.
2.1 Smoothing-Splines

Restricted to the Sobolev space $W$ of functions with second order weak derivatives, the solution to the problem of minimizing the functional

$$
\sum_{t=1}^{T} \omega_t(y_t - f(x_t))^2 + J_\lambda(f),
$$

where $(\omega_1, ..., \omega_T)'$ is a vector of weights and $J_\lambda(f) = \lambda \int f''(t)^2 dt$, is a cubic spline whose knots are the observations $x_t$, see Green and Silverman (1994). In particular, by taking $\omega_t = \frac{1}{\delta^2} E_\theta[\psi(\sigma_t)^{-1}|z_t]$, the estimate of $f$ that minimizes (16), with $\zeta$ and $\delta$ fixed, is the same one that maximizes (14). In this particular case, $B = [B_j(x_i)]_{j=1}^{T}$ with $B_1, ..., B_T$ being a set of natural cubic splines with knots $x_1, ..., x_T$. In the context of the EM algorithm, at its $k$th step, $f^{(k)} = B^{(k)}$, where

$$
e^{(k)} = (B^\top W_T B + 2\lambda(\delta^2)^{(k-1)}\Omega)^{-1}B^\top W_T y,
$$

and

$$
\Omega = \left[\int B''_i(t)B''_j(t)dt\right]_{i,j=1}^{T}.
$$

2.2 Penalized $B$-Splines

Smoothing splines is still computationally demanding, since the number of parameters is at least equal to the number of observations. A way to make the computational cost cheaper is to approximate $f$ via $B$-splines, i.e. by $f(x) \approx \sum_{j=1}^{M} c_j B_{m,j}(x)$, where the functions $B_{m,j}(x)$ are B-splines of order $m$ and to consider an estimate of $f$ of the form $\hat{f}(x) = \sum_{j=1}^{M} \hat{c}_j B_{m,j}(x)$. If $B = [B_{m,j}(x_i)]$ is the $(T \times M)$-matrix based on the B-splines basis calculated at $x_1, ..., x_T$, then $Be$ is an approximation to $(f(x_1), ..., f(x_T))^\top$. Replacing $f''(x)$ by $\hat{f}''(x)$ in $J_\lambda$, we get

$$
\Omega = \left[\int B''_{m,j}(t)B''_{m,k}(t)dt\right]_{j,k=1}^{M}.
$$

As before, at the $k$th step of the EM algorithm, we find $e^{(k)}$ as in (17) with $B$ and $\Omega$ as defined in this section.

2.3 $P$-Splines

A problem with the use of penalized $B$-splines as described above is the matrix $\Omega$ in (19). As anticipated in the introduction, its computation can be extremely complex, especially if we consider derivatives of a higher order as suggested by Wand and Ormerod (2008). An alternative is the use of $P$-splines, in which case the penalty $J_\lambda$ is given by $\lambda \sum_{j=k+1}^{M} (\Delta^k c_j)^2$. Here, $\Delta$ is just the difference operator $\Delta c_j = c_j - c_{j-1}$ and $k$ emulates in a sense the order of the derivative in $J_\lambda(f) = \lambda \int f''(t)^2 dt$, since the differences $\Delta^k$ are
simply a discrete approximation of the \( k \)th derivative. Actually, Eilers and Marx (1996) show that there is
a close connection between the penalties (19) and their proposal. Now, if we denote by \( D_k \in \mathbb{R}^{(M-k)\times M} \)
the difference matrix operator, so that \( D_k c \) is just the vector of differences of order \( k \) calculated from \( c \), then
\[
\sum_{j=k+1}^M (\Delta^k c_j)^2 = \|D_k c\|^2,
\]
so that \( \Omega = D_k^T D_k \).

**Algorithm 2.1** Estimation algorithm.
1. get the initial estimate \( f^{(0)} \) (a way to get the initial is assuming the errors i.i.d and Gaussian and using
penalized least squares, for example.);
2. get \( \delta^{(0)} \) and \( \zeta^{(0)} \);
3. do until convergence
   - **Step \((k + 1).1\)**: \( c^{(k+1)} = (B^T W_T(\theta^{(k)})B + 2\lambda(\delta^{(k)})\Omega)^{-1} B^T W_T(\theta^{(k)}) y; \)
   - **Step \((k + 1).2\)**: \( f^{(k+1)} = Bc^{(k+1)}; \)
   - **Step \((k + 1).3\)**: \( (\delta^{(k+1)}) = T^{-1} (y - f^{(k+1)})^T W_T(c^{(k+1)}, \delta^{(k)}, \zeta^{(k)})(y - f^{(k+1)}); \)
   - **Step \((k + 1).4\)**: get \( \zeta^{(k+1)} \) by solving the equation \( \frac{\partial}{\partial \zeta} C(\zeta; c^{(k+1)}, \delta^{(k+1)}, \zeta) = 0. \)

### 2.4 Confidence Bands and Smoothing Parameter Selection

A way to obtain confidence bands for the estimates of \( f, \delta \) and \( \zeta \) is through bootstrap. We describe below two
approaches, the parametric and nonparametric ones. The first one takes into account the distribution of the
errors (viz. a mixture of Normals), while the second one uses the empirical distribution of the residuals. In
practice, the nonparametric bootstrap is most often used.

For the parametric case, we use the fact that \( \epsilon_t|\sigma_t \sim \mathcal{N}(0, \psi(\sigma_t)^2) \) and \( \sigma_t \sim h \) to simulate \( \sigma^*_1, ..., \sigma^*_T \)
according to \( h \) and then, for each \( t = 1, ..., T \), obtain realizations \( \epsilon^*_t, ..., \epsilon^*_T \) from \( \mathcal{N}(0, \psi(\sigma^*_t)^2) \). The next
step consists in defining \( y^*_t = \hat{f}(x_t) + \tilde{\delta} \epsilon^*_t \), where \( \hat{f} \) and \( \tilde{\delta} \) are the estimates of \( f \) and \( \delta \), respectively, obtained
from the original sample \( y_1, ..., y_T \). The bootstrap sample is then used to get bootstrap estimates \( \hat{f}^*_j \) of \( f \).
This procedure is repeated \( G \) times to construct the desired confidence bands. In the nonparametric one \( \epsilon^*_t \) is
randomly sampled from the residuals without assuming any specific distribution.

**Remark.** As noted by Lahiri (2003) and references there in, if the error has infinite variance, both for regression
models and time series, usual bootstrap methods may fail if the condition \( G = o(T) \) does not hold. In
particular, care should be taken if we assume, for example, a stable distribution for the error.

The smoothing parameter \( \lambda \), on the other hand, directly influences the smoothness of the estimator \( \hat{f} \) in
penalized regression models and, therefore, its variance and bias as well. There is more than one way to
select such a tuning parameter, being one possibility the calculation of the AIC or BIC for different values of
The necessary modifications in Algorithm 2.1 are summarized in Algorithm 2.2. To the conditional density (11), and consider the following approximations of $W_{\sigma}$ Carlo methods to approximate it. For this, let $Q$ When the expectation 2.5 Approximation of $S(\theta|\theta')$ via Monte Carlo

When the expectation $Q(\theta|\theta^{(j)})$ cannot be computed analytically or it is difficult to do so, one can use Monte Carlo methods to approximate it. For this, let $\sigma_{t,1}^{(j)}, ..., \sigma_{t,\lambda}^{(j)}$ be a pseudo random sample from $\sigma_t$ according to the conditional density (11), and consider the following approximations of $W$ and $C(\zeta; \theta')$, respectively,

$$\hat{W}_j \equiv \text{diag}\left\{ \sum_{i=1}^{J(j)} \psi(\sigma_{1,i}^{(j)})^{-2}, ..., \sum_{i=1}^{J(j)} \psi(\sigma_{T,i}^{(j)})^{-2} \right\}.$$  \hspace{1cm} (20) $$

and $\hat{C}(\zeta; \theta^{(j)}) = \sum_{t=1}^{T} \sum_{i=1}^{J(j)} \log[h(\sigma_{t,i}^{(j)})/\psi(\sigma_{t,i}^{(j)})]$. Then $S(\theta|\theta^{(j)})$ can be approximated by

$$\hat{S}(\theta|\theta^{(j)}) = \hat{C}(\zeta; \theta^{(j)}) - T \log \delta - \frac{1}{2} (y - Bc)^\top \hat{W}_j (y - Bc) - \lambda c^\top \Omega c.$$  \hspace{1cm} (21) $$

The necessary modifications in Algorithm 2.1 are summarized in Algorithm 2.2.

**Algorithm 2.2 Monte-Carlo steps**

**Step 1.** Keep $\zeta = \zeta^{(j)}$ and get $c^{(j+1)}$ by finding the solution of $(B^\top \hat{W}_j B + 2\lambda \delta^2 \Omega)c = B^\top \hat{W}_j y$.

**Step 2.** Use $c^{(j+1)}$ to get $\delta^{(j+1)} = T^{-1}(y - Bc^{(j+1)})^\top \hat{W}_j (y - Bc^{(j+1)})$.

**Step 3.** Use $c^{(j+1)}$ and $\delta^{(j+1)}$ to estimate $\zeta^{(j+1)}$ by maximizing $\hat{S}(\theta|f^{(j+1)}, \delta^{(j+1)}, \zeta^{(j)})$.

**Remark.** If $\psi$ does not depend on $\zeta$, then in the 3rd step of Algorithm 2.2, $\zeta$ is updated simply by solving the equation $\frac{\partial}{\partial \zeta} \hat{C}(\zeta; \theta^{(j)}) = 0$.

**Remark.** The vector $\zeta = (\zeta_1, ..., \zeta_p)^\top$ usually has small dimension and the maximization with respect to the
elements of this vector can be broken in a small sequence of univariate maximizations,

\[
\begin{align*}
\zeta_1^{(j+1)} &= \arg\max_{\zeta_1} \tilde{S}(f^{(j+1)}, \delta^{(j+1)}, \zeta_2^{(j)}, \ldots, \zeta_p^{(j)}) \\
& \vdots \\
\zeta_p^{(j+1)} &= \arg\max_{\zeta_p} \tilde{S}(f^{(j+1)}, \delta^{(j+1)}, \zeta_1^{(j+1)}, \zeta_2^{(j+1)}, \ldots, \zeta_p).
\end{align*}
\]  

(22)

In the Algorithm 2.3 we describe, in a more concise way, the steps of the estimation procedure of the target function \( f \) and of the vector of parameter \( \zeta \). If the expectations of the step E of the algorithm and the conditional distribution of \( \sigma \), given \( y \), can be computed analytically, the steps 1 to 5 can be replaced by the usual steps E and M of the algorithm, i.e. without MC simulations.

**Algorithm 2.3** Generic estimation algorithm for the semiparametric model.

1. get the initial estimate \( f^{(0)} \). (a way to get the initial is assuming the errors i.i.d and Gaussian and using penalized least squares, for example.)
2. get \( \zeta^{(0)} \)
3. do until convergence

   \( k \)th iteration: given \( f^{(k-1)} \), \( \delta^{(k-1)} \) and \( \zeta^{(k-1)} \):
   
   **Step k.1:** for each \( t = 1, \ldots, T \) generate a pseudo random sample \( \sigma^{(1)}_t, \ldots, \sigma^{(J(j))}_t \) from

   \[
   \tilde{k}(\sigma_t | y_t, \theta^{(k-1)}) \propto \frac{1}{\delta^{(k-1)} \psi_{\theta^{(k-1)}}(\sigma_t)} \phi \left( \frac{y_t - f_t^{(k-1)}}{\delta^{(k-1)} \psi_{\theta^{(k-1)}}(\sigma_t)} \right) h(\sigma_t | \theta^{(k-1)})
   \]

   where \( f_t^{(k-1)} \) is the approximation obtained for \( f \) at the value \( x_t \) in iteration \( (k - 1) \).
   
   **Step k.2:** get \( \tilde{W}_k \) according to (20);
   
   **Step k.3:** get \( f^{(k)} \) through the appropriate normal equations;
   
   **Step k.4:** get \( \zeta^{(k)} \).

2.5.1 On the implementation of the Metropolis-Hastings algorithm

To sample from the conditional density of \( \sigma \), we can take as the proposed density the marginal \( h \) itself, i.e. \( q(\sigma_t^{(j)}, s) \equiv h(s | \theta) \), where \( \theta \) is the vector of parameters estimated in the last iteration of the algorithm. In this
case, the acceptance probability \( \rho(y_t, f(x_t)) \) is given by

\[
\rho(y_t, f(x_t)) = \min \left\{ \frac{\psi(y_t)}{\psi(s)} \exp \left[ - \frac{(y_t - f(x_t))^2}{2(\delta_y)^2} \left( \frac{1}{\psi(s)^2} - \frac{1}{\psi(y_t)^2} \right) \right], 1 \right\},
\]

so that the sample of interest is generated according to the rule \( y_{t+1} = s \), with probability \( \rho(y_t, f(x_t)) \) and equal to \( y_t \) with probability \( 1 - \rho(y_t, f(x_t)) \).

For example, in the case of the Student’s t distribution with \( \nu \) known, the acceptance probability becomes

\[
\rho(y_t, f(x_t)) = \min \left\{ \sqrt{\frac{s}{y_t}} \exp \left[ - \frac{(y_t - f(x_t))^2}{2\zeta^2} \left( s - y_t \right) \right], 1 \right\}.
\]

This algorithm can also be used for models with errors distributed according to a stable distribution, since we just need to be able to generate a pseudo random sample from this distribution, avoiding the difficulties of maximizing the corresponding likelihood function.

### 2.6 Extension to Multivariate Models

The above methodology is easily extended to situations with more than one covariate. To illustrate, we will consider here two very common cases: the partially linear models and the additive models. Partially linear models are of the form

\[
y_t = \beta^\top z_t + f(x_t) + \delta \epsilon_t,
\]

where \( z_t = (z_{t1}, ..., z_{tp})^\top \) is a vector of covariates associated to the linear part of the model. The other components are just as before. Additive models, on the other hand, assume that

\[
y_t = \mu + \sum_{i=1}^p f_i(x_{it}) + \delta \epsilon_t,
\]

where \( f_1, ..., f_p \) are unknown functions in some suitable function space. For instance, when dealing with splines, it is usual to consider Sobolev spaces. In both cases, estimation are very similar to the univariate case. Only a couple of remarks are necessary. First, specifically for additive models, the penalization term must take into account all \( p \) functions. In our case, where the errors are assumed to follow a scale mixture of Normals, the objective function is then given by

\[
\frac{1}{T} \sum_{t=1}^T \ell(f_1, ..., f_p, \delta, \zeta; y_t) + \sum_{i=1}^p J_{i,\lambda_i}(f_i).
\]

Again, common choices for the penalty functionals are \( J_{i,\lambda_i}(f_i) = \lambda_i \int f_i''(s)^2 \, ds \) (smoothing splines) or, if we approximate \( f_i \) by B-splines, so that \( f(x) \approx \sum_{j=1}^M c_j B_{m,j}(x) \), then we could consider \( J_{i,\lambda_i}(c_i) = \lambda_i c_i^\top \Omega c_i \),
where $\Omega$ is chosen according to the specific estimation methodology (e.g. P-splines). Secondly, still regarding additive models, it is necessary to restrict the target functions in some way in order to identify the model. Here, we assume the usual conditions $\sum_{t=1}^{T} f_i(x_{ti}) = 0$, for all $i = 1, \ldots, p$. In the partially linear model, the identification issue is circumvented by not allowing an intercept in the linear component. The last remark is about the estimation algorithm for both partially linear and additive models. The standard approach is based on the backfitting algorithm, which simply consists in updating in the EM algorithm the estimates related to each predictor by letting the current estimates related to the other predictors fixed (Hastie and Tibshirani, 2009). One can find the general estimation routine in Algorithms 2.4 and 2.5. The additive model is illustrated in the simulation study as well as in the application.

### Algorithm 2.4 Estimation algorithm for the additive model.

1. set $f_1^{(0)} = \ldots = f_p^{(0)} = 0$;
2. get $\delta^{(0)}$ and $\zeta^{(0)}$;
3. $\hat{\alpha} = \frac{1}{T} \sum_{t=1}^{T} y_t$.
4. do until convergence

   For each $i = 1, \ldots, p$,
   
   **Step** $(k+1).0$: Let $y^{(i)} = y - \hat{\alpha}1 - \sum_{i_0 \neq i} B_{i_0} \hat{c}_{i_0}$
   
   **Step** $(k+1).1$: $c_i^{(k+1)} = (B_i^TW_T(\theta^{(k)})B_i + 2\lambda_i(\hat{\delta}^2)^{(k)}\Omega)^{-1}B_i^TW_T(\theta^{(k)})y^{(i)}$;
   
   **Step** $(k+1).2$: $f_i^{(k+1)} = B_i c_i^{(k+1)}$;

   **Step** $(k+1).3$: Set $\hat{y}^{(k+1)} = \hat{\alpha}1 + \sum_{i=1}^{p} f_i^{(k+1)}$ and get
   
   $(\hat{\delta}^2)^{(k+1)} = T^{-1}(y - \hat{y}^{(k+1)})^TW_T(c^{(k+1)}, \delta^{(k)}, \zeta^{(k)})(y - \hat{y}^{(k+1)})$;

   **Step** $(k+1).4$: get $\zeta^{(k+1)}$ by solving the equation $\frac{\partial}{\partial \zeta} C(\zeta; c^{(k+1)}, \delta^{(k)}, \zeta) = 0$.

Note: $\hat{c}_i$ stands for the last estimate update of $c_i$.

### 3 Simulation Studies

For the sake of illustration, we consider the target functions $f_1(x) = \sin(20(x + 0.2))/(x + 0.2)$ and $f_2(x) = 10x + 15e^{-12x^2}$ in the following simulation studies.
Algorithm 2.5 Estimation algorithm for the partially linear model.

1. get the initial estimates $f^{(0)}$, $\beta^{(0)}$, $\delta^{(0)}$ and $\zeta^{(0)}$;

2. do until convergence

   Step $(k+1).1$: $c^{(k+1)} = (B^\top W_T(\theta^{(k)})B + 2\lambda(\delta^2)^{(k)} \Omega)^{-1}B^\top W_T(\theta^{(k)})(y - Z\beta^{(k)})$;

   Step $(k+1).2$: $f^{(k+1)} = Bc^{(k+1)}$;

   Step $(k+1).3$: $\beta^{(k+1)} = (Z^\top W_T(\theta^{(k)})Z)^{-1}Z^\top W_T(\theta^{(k)})(y - Bc^{(k+1)})$;

   Step $(k+1).4$: Set $\hat{y}^{(k+1)} = Z\beta^{(k+1)} + f^{(k+1)}$ and get

   $$(\delta^2)^{(k+1)} = T^{-1}(y - \hat{y}^{(k+1)})^\top W_T(c^{(k+1)}, \delta^{(k)}, \zeta^{(k)})(y - \hat{y}^{(k+1)});$$

   Step $(k+1).5$: get $\zeta^{(k+1)}$ by solving the equation $\frac{\partial}{\partial \zeta}C(\zeta; c^{(k+1)}, \delta^{(k+1)}, \zeta) = 0$.

---

(a) Target function $f_1$  
(b) Target function $f_2$

Figure 1: Target functions used in the simulation studies for the semiparametric regression model.

3.1 Estimation via B-Splines

We start considering the target function $f_1(x)$ and scale parameter $\delta = 1.5$. For the simulated data, we assumed $\varepsilon_t \sim t_{\nu_o}$, with $\nu_o = 1.5$, and generate a sample of size 200. Three fits were done, and for each one we took $M = 60$ and smoothing parameters chosen via cross-validation, $\lambda = \lambda_{cv}$. To assess the 95%-confidence bands associated to $f$ and the 95%-confidence interval associated to $\delta$, we generated $G = 1000$ bootstrap samples. All fits were done with the proposed methodology with the P-splines penalization type.

For the first fit (Fit 1) we assumed $\varepsilon_t \sim t_{\nu}$, with $\nu = \nu_o = 1.5$ and smoothing parameter $\lambda_{cv} = 8.447$. For the second fit (Fit 2) we considered again the Student’s $t$ distribution, but with $\nu = 3.5$ and $\lambda_{cv} = 2.710$. Finally, in the third fit (Fit 3), we estimated $f$ via penalized least squares ($\lambda_{cv} = 79.045$). The results are shown in Table 1 and Figure 2. The estimated curve is shown in full line and confidence bands in dashed lines. As one can see, the figures show how better the estimators based on the true distribution perform quite better than the simple penalized least squares procedure in the presence of outliers (in this case, produced by an extremely
Table 1: Estimates and confidence intervals for the scale parameter $\delta = 1.5$ as well as the mean square errors (MSE) and $R^2$ for each fit in the first simulation study.

<table>
<thead>
<tr>
<th>Fit</th>
<th>bootstrap</th>
<th>$\tilde{\delta}$</th>
<th>CI$_{95%}$</th>
<th>MSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>parametric</td>
<td>1.374</td>
<td>[1.084, 1.584]</td>
<td>0.1785</td>
<td>0.9248</td>
</tr>
<tr>
<td></td>
<td>nonparametric</td>
<td></td>
<td>[1.062, 1.615]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>parametric</td>
<td>1.939</td>
<td>[1.610, 2.133]</td>
<td>0.1674</td>
<td>0.9294</td>
</tr>
<tr>
<td></td>
<td>nonparametric</td>
<td></td>
<td>[1.483, 2.346]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 (LS)</td>
<td>parametric</td>
<td>5.317</td>
<td>[4.611, 5.515]</td>
<td>1.8044</td>
<td>0.2401</td>
</tr>
<tr>
<td></td>
<td>nonparametric</td>
<td></td>
<td>[3.420, 7.023]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

heavy tailed distribution). This remained true even when we used a larger degree of freedom (3.5 instead of 1.5) to run the penalized likelihood procedure. The confidence bands confirm this pattern. It is worth noticing that the outliers excessively pull (or push) the least squares estimates away from the true function, so that the latter stays out of the confidence region at some places. We get the same impression when we compare the respective MSE and $R^2$. Here, the $R^2$ stands for the proportional decrease in model error and is defined by $R^2 = (MSE_o - MSE)/MSE_o$, where $MSE_o = T^{-1} \sum_{i=1}^{n}(\bar{y} - f(x_i))^2$, so that values close to one indicate a fit far better than simply averaging the response variable. When we compare our results, we see an $R^2 \approx 0.92$ for both Fit 1 and 2, and $R^2 \approx 0.24$ for the penalized least squares fit. Regarding the scale parameter, we find similar results: we see no significant differences if we choose a larger (but $\neq \nu_o$) value of $\nu$, but a substantial difference when we compare with the penalized least squares estimates. It is worth noting that in this last case, the outliers induce both a much larger smoothing parameter to control variance and a much larger scale parameter. In particular, when the error distribution is correctly specified, we get the estimate $\tilde{\delta} = 1.374$ with coverage probability equal to 87.3% when using parametric bootstrap and 89.2% when using nonparametric bootstrap. We have not calculated coverage probability for the scale parameter estimates with misspecified distributions because, especially those with tails not so heavy, the presence of extreme values tends to enlarge the estimate of $\delta$ (in comparison to the true value under the correct model) to accommodate them.

In the second simulation study, we empirically assess the consistency of the proposed estimator using P-splines and its sensitivity misspecified distributions. By considering the target function $f_2$ and scale parameter $\delta = 1.5$, we simulated 1,000 samples of different sizes ($T = 50, 100, 200, 500, 1000$) based on the Cauchy distribution. In each adjustment, we used $M = 60$ and a smoothing parameter chosen via cross-validation for $T = 500$ for each distribution as a proxy. More specifically, we fit the data by taking into account the Cauchy, Student’s t with $\nu = 7$, Laplace and Normal distributions. The average curves obtained and their respective 95% bands are found in Figure 3. The figure suggests that from the target function estimation perspective, all methods based on heavy tailed distributions seem asymptotically unbiased estimates with decreasing variance as we increase the sample size. On the other hand, the estimates based on the Normal distribution did not perform well. It is worth remembering that the model error follows the Cauchy distribution, which has infinite variance. This fact, in itself, invalidates most asymptotic results for regression models based on penalized
The estimates are quite sensitive to extreme values and, as we increase $T$, new outliers arise, disturbing the corresponding estimates. Table 2, on the hand, shows the evolution of the average MSE and $R^2$ as $T$ increases when we use the Cauchy distribution. In it, we also observe the behavior of the $\delta$ estimator for different sample sizes. For the sake of comparison, we included the MSE results for the other heavy tailed distributions as well. Finally, as one can see, the estimates of $\delta$ also seem to converge to the true value 1.5 when we use the Cauchy distribution. Moreover, the corresponding standard deviation also reduces consistently toward zero when we increase the sample size.

To conclude this section, we present a brief illustration of the approach described in Algorithm 2.4 for additive models. We considered here a sample of size 100 obtained from the model $y_t = f_1(x_1) + f_2(x_2) + 1.5\varepsilon_t$, with $\varepsilon_t \sim t_{3.0}$. We ran 1,000 simulations and in each one estimated all the corresponding parameters. Fig-

Figure 2: Estimates of $f_1$ using the proposed methodology for different distributions.
Figure 3: Average estimates of $f_2$ (over 5000 simulations) using penalized splines for different scale mixtures of Normals and their respective 95% confidence ranges.

Figures 4b and 4a show the average curves of the target functions $f_1$ and $f_2$ and their respective 95%-confidence bands and in Figure 4c it is found the combined view of these estimates. The average of the estimated scale parameter is 1.35 with standard deviation 0.11, while the model average MSE was 0.37 and standard deviation 0.13.
Table 2: Estimates of scale parameter $\delta$ (only for the Cauchy distribution) and metrics (expected) MSE and $R^2$ for different sample sizes. The values in parentheses represent standard deviations.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\delta$</th>
<th>$R^2$</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cauchy</td>
<td>0.98 (0.33)</td>
<td>0.894 (0.099)</td>
<td>1.40 (1.00)</td>
</tr>
<tr>
<td></td>
<td>1.26 (0.22)</td>
<td>0.956 (0.037)</td>
<td>0.57 (0.34)</td>
</tr>
<tr>
<td></td>
<td>1.37 (0.15)</td>
<td>0.978 (0.016)</td>
<td>0.28 (0.13)</td>
</tr>
<tr>
<td></td>
<td>1.43 (0.10)</td>
<td>0.9885 (0.0082)</td>
<td>0.150 (0.062)</td>
</tr>
<tr>
<td></td>
<td>1.466 (0.068)</td>
<td>0.9946 (0.0034)</td>
<td>0.069 (0.024)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Cauchy</td>
<td>0.98 (0.33)</td>
</tr>
<tr>
<td></td>
<td>0.894 (0.099)</td>
</tr>
<tr>
<td></td>
<td>1.40 (1.00)</td>
</tr>
<tr>
<td>Laplace</td>
<td>1.27 (1.03)</td>
</tr>
<tr>
<td></td>
<td>2.66 (2.34)</td>
</tr>
</tbody>
</table>

Figure 4: Curve estimates for the additive model

3.2 Real Data

To illustrate the above methodology, we consider data used by Harrison and Rubinfeld (1978) in a study on the use of real estate market price data to assess the willingness to pay for improvements in air quality. Although
they used a larger collection of variables, we considered here only units two of them here: the proportion of owners of units built before 1940 \((x)\) and the levels of nitrogen oxide concentration in pphm \((y)\) as a proxy of air pollution. We took a sample of size \(T = 150\). The data can be seen in Figure 5 together with three different estimates of the function relating the variables of interest. Visual inspection of the data suggests the such a relationship is contaminated with heavy tailed errors. An indication of this is the 95\%-confidence bands obtained in each adjustment, as seen by comparing the adjustments assuming the Laplace distribution (Figure 5b) and the Normal distribution (Figure 5c). The same, however, could not be said if we compare the confidence bands associated to the Normal distribution with the ones associated to the Student’s t distribution (Figure 5a). In order to select a model for the data, we compared several distributions by evaluating the Kolmogorov-Smirnov (KS) statistic applied to the residues standardized by the scale parameter. In each case, the comparison distribution was the one used to fit the data and even with an apparently less smooth estimate, the model with the lowest KS statistic was the one assuming \(\varepsilon_t \sim t_{4.0} (KS = 0.0664)\). For comparison, the models assuming the Laplace (Figure 5b) and Normal distributions resulted, respectively, in \(KS = 0.4304\) and \(KS = 0.1111\) (this one with p-value 0.049). The scale parameter estimate for the selected model \((\varepsilon_t \sim t_{4.0})\) was \(\hat{\delta} = 0.049\) with a 95\%-confidence interval \([0.039; 0.055]\). For the sake of comparison, when considering the Laplace distribution we get \(\hat{\delta} = 0.0042\) \([0.0033; 0.0043]\), and for the Normal distribution, \(\hat{\delta} = 0.073\) \([0.064; 0.080]\).

![Figure 5](image-url)  
(a) \(\varepsilon_t \sim t_{4.0}\)  
(b) \(\varepsilon_t \sim \text{Laplace}\)  
(c) \(\varepsilon_t \sim N(0, 1)\)

Figure 5: Estimates for the real data set and the corresponding 95\%-confidence bands.

### 3.3 Other Approximation Methods

The next two simulation studies illustrate how well the proposed methods adequates to other bases functions, namely the Fourier basis and the Haar wavelet basis.

#### 3.3.1 Trigonometric Series

For this simulation, we consider again the target function \(f_1\) and scale parameter \(\delta = 1.5\). To assess the estimation performance, we generated 5,000 samples of size 500 according to a Student’s t distribution with
\( \nu = 1.5 \). The function \( f_1 \) and the scale parameter were estimated assuming the Student’s \( t \) distribution with degrees of freedom 1.5, 5.0 and 15.0 as well as using least squares. The basis functions were based on trigonometric functions, so that \( \tilde{f}(x) = c_1 + \sqrt{2} \sum_{j=2}^{M} c_j \cos((j - 1)\pi x) \). In this case, \( M \) stands for the smoothing parameter, so that \( \lambda = 0 \) in Algorithm 2.1 and \( (T \times M) \) design matrix is \( B = [b_{ij}] \), with \( b_{11} = 1 \) and \( b_{ij} = \cos((j - 1)\pi x_i) \), para \( j \geq 2 \). The proxy \( M = 9 \) was empirically chosen by comparing several scenarios. The average curves obtained and their respective 95% bands are found in Figure 6. As expected, the estimated curves assuming a Student’s \( t \) fit almost perfectly to the real curve (dotted), while the estimate using least squares (Figure 8d) is extremely sensitive to extreme values. For the sake of comparison, we present the respective average MSE’s and \( R^2 \) in Table 3 together with the corresponding estimates of the scale parameter. As can be seen, estimates based on Student’s \( t \) distribution have a performance far superior to that

![Figure 6: Average estimates of \( f_1 \) (over 5000 simulations) using trigonometric series for different scale mixtures of Normals and their respective 95% confidence ranges.](image_url)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>( \delta )</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_{1.5} )</td>
<td>1.486 (0.085)</td>
<td>0.095 (0.036)</td>
</tr>
<tr>
<td>( t_{5.0} )</td>
<td>2.32 (0.15)</td>
<td>0.117 (0.045)</td>
</tr>
<tr>
<td>( t_{15.0} )</td>
<td>3.24 (0.29)</td>
<td>0.169 (0.076)</td>
</tr>
<tr>
<td>Least Squares</td>
<td>11.5 (25.8)</td>
<td>13.9 (235.2)</td>
</tr>
</tbody>
</table>
obtained using least squares. The high standard deviations observed in this latter case are due to the presence of extreme values and indicate that in many of the performed simulations the corresponding observed MSE was much higher than the average value indicated in the table. The same does not happen in the other cases.

3.3.2 Wavelet expansion

In this section we use the fact that any function \( f \in L^2[0,1] \) can be represented by a wavelet series

\[
f(x) = \sum_{k=0}^{\infty} c_{j_0,k} \phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} d_{j,k} \psi_{j,k}(x),
\]

where \( j_0 \) is a coarser scale, \( \phi_{j,k} \) and \( \psi_{j,k} \) are derived from the father (scaling function) and mother wavelet functions \( \phi \) and \( \psi \), respectively, through dilations and translations, \( j, k \in \mathbb{Z} \), and where \( c_{j,k} \) and \( d_{j,k} \) are the wavelet coefficients, obtained by

\[
c_{j,k} = \int f(x) \phi_{j,k}(x) \, dx \quad \text{and} \quad d_{j,k} = \int f(x) \psi_{j,k}(x) \, dx,
\]

respectively. We consider a special wavelet basis, namely the Haar basis, and use in this case the expansion (with \( j_0 = 0 \))

\[
f(x) = c_{0,0,0}(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} d_{j,k} \psi_{j,k}(x),
\]

see Vidakovic (1999). In Fig. 7 we show multiresolution analyses (MRA) corresponding to the signal

\[
y_t = \sin(10t) + \epsilon_t,
\]

where \( \epsilon_t \sim N(0, 1) \), 7a, and \( \epsilon_t \sim t(2) \), Figure 7b. In the Gaussian case, the larger wavelet coefficients correspond to the signal, while in the case of a t distribution, the larger coefficients correspond to noise. Therefore, the application of the usual universal threshold would be not effective in the second case. We recall that the expression of the universal threshold (see Donoho and Johnstone, 1994) is given by

\[
\lambda_U = \delta \sqrt{2 \log T},
\]

where \( \delta \) is the scale parameter (standard deviation in the case of Gaussian noise). Since the Student’s t distribution is a scale mixture of Normals, we can obtain an “equivalent universal threshold”, for signals with errors following a Student’s \( t_\nu \) distribution. Indeed, as indicated in the Appendix, it should be proportional to

\[
\sqrt{T^{\nu/2} - 1}.
\]

(24)

For the simulation study, given that wavelets are extremely efficient in capturing discontinuities of the target function, we chose the following target function:

\[
f_3(x) = \begin{cases} 
0, & \text{if } 0 \leq x < \frac{1}{4} \\
2, & \text{if } \frac{1}{4} \leq x < \frac{1}{2} \\
-3, & \text{if } \frac{1}{2} \leq x < \frac{3}{4} \\
6, & \text{if } \frac{3}{4} \leq x < 1 
\end{cases}
\]

(25)
We supposed $\epsilon_t$ to follow a Cauchy distribution and a scale parameter $\delta = 1.5$. These being fixed, we generated a pseudo random sample of size $T = 512$ (a power of 2, to make easier the application of a fast wavelet transform) according to the model $y_t = f(x_t) + \delta \epsilon_t$ where the $x_t$’s are equally spaced in $[0, 1]$.

The simulation configuration in this last exercise is essentially the same as the previous one, but with the target function $f_3$ an using the Haar basis instead of trigonometric series. The estimation results are found in Figure 8. As for the scale parameter and average MSE, the results are in Table 4. The conclusions here are essentially the same as in the case of trigonometric series.

Table 4: Scale parameter estimates as well as the average MSE for different distributions using Haar wavelets. The values in parentheses stand for standard deviations.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\delta$</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{1.5}$</td>
<td>1.494 (0.086)</td>
<td>0.032 (0.023)</td>
</tr>
<tr>
<td>$t_{5.0}$</td>
<td>2.38 (0.15)</td>
<td>0.041 (0.029)</td>
</tr>
<tr>
<td>$t_{15.0}$</td>
<td>3.25 (0.29)</td>
<td>0.066 (0.048)</td>
</tr>
<tr>
<td>Least Squares</td>
<td>11.5 (26.0)</td>
<td>6.5 (112.9)</td>
</tr>
</tbody>
</table>
Figure 8: Average estimates of $f_2$ (over 5000 simulations) using wavelet (Haar) series for different scale mixtures of Normals and their respective 95% confidence ranges.

4 Conclusion

In this paper we have shown that the use of distributions defined as mixtures of Gaussian distributions in the scale results in robust estimators of the unknown function in semiparametric models. We also consider the use of the EM algorithm to obtain maximum likelihood estimate together with other techniques of function approximation, like splines an wavelets. Through several simulation studies we have shown that the proposed technique works well, consistently better than least squares, for example. Given the nature of the distributions entertained, there is a close connection between the classical estimation procedure and the Bayesian one. This will be the object of other research, see Taddeo and Morettin (2019).

Acknowledgements. The authors are grateful for the partial support from grants of CNPq (MMT) and FAPESP 2018/04654-9(PAM).

References


Appendix

The multivariate \( t \) distribution

We say that the random vector \( Z = (Z_0, ..., Z_{T-1})' \) follows a multivariate \( t \) distributions with \( \nu \) degrees of freedom, mean \( \mu \) an \( d \) correlation matrix \( R \) (or \( \Sigma \) as covariance matrix) if

\[
f(z) = \frac{\Gamma\left(\frac{\nu+T}{2}\right)}{(\pi \nu)^{T/2} \Gamma\left(\frac{\nu}{2}\right) |R|^{1/2}} \left[ 1 + \frac{1}{\nu}(z - \mu)'R^{-1}(z - \mu) \right]^{-\frac{\nu+T}{2}}.
\]

If \( \mu = 0 \), we say that the distribution is central and, in the general case, use the notation \( Z \sim t_{\nu,T}(\mu, R) \).

Moreover, \((Y, S)\) follows a multivariate normal-gamma distribution with parameters \( \mu, R, \nu \) and \( \tau \), if

\[
\begin{align*}
Y | S = s & \sim N_T\left(\mu, \frac{R}{s}\right), \\
S & \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu \tau}{2}\right).
\end{align*}
\]

Then, \( Y \sim t_{\nu,T}(\mu, \sigma R) \). In particular, if \( \sigma = 1 \), then \( Y \sim t_{\nu,T}(\mu, R) \).

Borel inequality

This inequality is used below. See Adler (1990).

**Theorem.** Let \( Z = \{Z_t\}_{t \in T} \) be a Gaussian process, with mean zero and bounded trajectories a.s. Define \( \|Z\|_1 \equiv \sup_{t \in T} |Z_t| \). Then, \( E\|Z\|_1 < \infty \) and, for all \( \lambda > 0 \),

\[
P\{\|Z\|_1 - E\|Z\|_1 > \lambda \} \leq 2e^{-\frac{\lambda^2}{2\sigma_T^2}},
\]

where \( \sigma_T^2 \equiv \sup_{t \in T} E|Z_t|^2 \).
Derivation of the threshold \( (33) \)

Let us denote
\[
 Z = S^{-1} Y,
\]
where
\[
\begin{cases}
  Y \sim \mathcal{N}_T(0, \Sigma), \\
  \chi \equiv \frac{\nu S^2}{\sigma^2} \sim \chi^2_{\nu},
\end{cases}
\]
such that, assuming that \( R \) is the correlation matrix of \( Y \), we will have \( Z \sim t_{\nu,T}(0, R) \). Suppose further \( E\|Y\|_1 = 0 \). Now, by the Borel inequality applied to the random vector \( Y \) and defining \( \Lambda \equiv \{ z : \|z\|_1 > \lambda \} \),
\[
P\{\|Z\|_1 > \lambda \} = P\{\Lambda\} = EI(\Lambda)
\]
\[
= EE[I(\Lambda)|S] = EP\{\|Y\|_1 > S\lambda|S\}
\]
\[
\leq 2Ee^{-\frac{\lambda^2 S^2}{2\sigma^2 T}} \equiv 2Ee^{\kappa \chi},
\]
where \( I \) denotes the indicator function, \( \kappa \equiv -\frac{\lambda^2 S^2}{2\sigma^2 T} < 0 \) and \( \chi \sim \chi^2_{\nu} \), with \( \sigma^2_T \equiv \sup_{0\leq t\leq T} EY^2_t \). Then, using the fact that
\[
Ee^{\kappa \chi} = \left( \frac{1}{1 - 2\kappa} \right)^{\nu/2} = \left( \frac{\sigma^2_T \nu}{\sigma^2_T \nu + \lambda^2 \sigma^2} \right)^{\nu/2} = \left( \frac{\nu^2}{\nu^2 + (\nu - 2)\lambda^2} \right)^{\nu/2},
\]
since \( \text{Var} Y_i = \frac{\sigma^2 Y}{\nu - 2} \), we get
\[
P\{\|Z\|_1 > \lambda \} \leq 2 \left( \frac{\nu^2}{\nu^2 + (\nu - 2)\lambda^2} \right)^{\nu/2}.
\]
Consequently, in order to obtain a rate of convergence of the probability \( (26) \) equivalent to the Gaussian case, the threshold should taken as
\[
\lambda_{\nu,T} = \frac{\nu}{\sqrt{\nu - 2} \sqrt{T^{\nu/2} - 1}}.
\]
Note that, for \( T \) fixed, \( \lim_{\nu \to 2} \lambda_{\nu,T} = \infty \).