Ann Inst Stat Math (2016) 68:589–619
DOI 10.1007/s10463-015-0506-0

Influence diagnostics for robust P-splines using scale mixture of normal distributions

Felipe Osorio

Received: 25 February 2014 / Revised: 9 January 2015 / Published online: 12 February 2015

Abstract It has been well documented that the presence of outliers and/or extreme data can strongly affect smoothing via splines. This work proposes an alternative for accommodating outliers in penalized splines considering the maximum penalized likelihood estimation under the class of scale mixture of normal distributions. This family of distributions has been an interesting alternative to produce robust estimates, keeping the elegancy and simplicity of the maximum likelihood theory. The aim of this paper is to apply a variant of the EM algorithm for computing efficiently the penalized maximum likelihood estimates in the context of penalized splines. To highlight some aspects of the robustness of the proposed penalized estimators we consider the assessment of influential observations through case deletion and local influence methods. Numerical experiments were carried out to illustrate the good performance of the proposed technique.

Keywords Cook distance · Local influence · Penalized EM algorithm · Scale mixtures of normal distributions

1 Introduction

Regression methods using splines are very attractive because they represent a flexible approach to fitting curves and are often used to find the underlying tendencies in the

Electronic supplementary material The online version of this article (doi:10.1007/s10463-015-0506-0) contains supplementary material, which is available to authorized users.

F. Osorio (✉)
Departamento de Matemática, Universidad Técnica Federico Santa María,
Av. España 1680, Valparaíso, Chile
e-mail: felipe.osorios@usm.cl

data. Discussions about smoothing and nonparametric regression can be found, for example, in Silverman (1985), Eilers and Marx (1996) and Ruppert et al. (2003).

The development of robust methodologies with the objective of attenuating the effect of outliers and/or influential observations in semiparametric regression has received considerable attention since the seminal works of Huber (1979) and Utreras (1981). They introduced the robust smoothing considering M-estimators. Their ideas have since been refined and applied to more general contexts. For example, Koenker et al. (1994) proposed the quantile smoothing splines, Oh et al. (2004) proposed M-estimation for smoothing periodic functions, while Lee and Oh (2007) and Oh et al. (2008) developed robust M-type estimation procedures with applications to additive mixed models and local polynomial regression, respectively. More recently, Tharmaratnam et al. (2010) and Mateos and Giannakis (2012) have discussed very efficient methods for computing penalized S- and M-type regression splines estimators, respectively. The appropriate selection of the smoothing parameter is crucial in the class of penalized spline regression models. It is important to note that this can also be strongly affected by the presence of outlying observations. To avoid this type of difficulty, Cantoni and Ronchetti (2001), Wei (2005) and Lee and Cox (2009) have focused on developing methods of robust selection of the smoothing parameter, while Staudenmayer et al. (2009) and Ibacache-Pulgar and Paula (2011) have described approaches for accommodating outliers in semiparametric regressions considering Student t errors.

With the objective of evaluating the model assumptions and determining whether outlying or extreme observations can influence the parameter estimates, diagnostic procedures have been developed in the context of semiparametric regressions (Eubank 1985). For example, Eubank (1984) studied the properties of the prediction matrix for smoothing splines, while Silverman (1985) discussed definitions for the residuals. Eubank and Gunst (1986) proposed measures to evaluate influence in penalized least squares that are useful in contexts like smoothing spline and ridge regression (Hoerl and Kennard 1970). Studies of influence in the context of ridge regression suggest that this class of penalized estimators can be very sensitive to extreme observations (Walker and Birch 1988; Billor and Loynes 1999; Shi and Wang 1999). Thomas (1991) studied influence to evaluate the impact of extreme observations on the selection of the smoothing parameter considering the local influence procedure proposed by Cook (1986). Manchester (1996) proposed a graphic tool to evaluate the sensitivity of some robust smoothing methods through the use of the influence function. Kim (1996) and Wei (2004) developed diagnostic measures in smoothing splines based on case deletion procedures. Kim et al. (2002) and Ibacache-Pulgar and Paula (2011) discussed influence diagnostics using elimination of observations and local influence, respectively, in partially linear models.

This work proposes an alternative to accommodate outliers in penalized splines, also known as P-splines (see Eilers and Marx (1996)), considering distributions with heavier tails than the normal. Specifically, we considered the class of scale mixtures of normal distributions (SMN), which includes as particular cases exponential power, contaminated normal, slash and Student t distributions, among others (Andrews and Mallows 1974). SMN distributions have often been proposed for developing robust inferences in various statistical models. This class of distribution inherits many of the

basic properties of the normal distribution and allows for maintaining the elegance and optimality of estimating parameters considering the maximum likelihood method (ML) under normality (Lange and Sinsheimer 1993; Jamshidian 1999). In this work we apply a penalized EM algorithm (Green 1990) to estimation in P-splines. An interesting characteristic of the proposed procedure is that the estimator of the coefficients adopts the form of a weighted smoother. The estimation procedure proposed in this work has been implemented in the heavy package (Osorio 2014), developed as an extension of the R statistical software (R Core Team 2014). The package is available from CRAN and the web site http://heavy.mat.utfsm.cl. We studied influence diagnostics to determine the robustness of the proposed procedure against outlying observations and some common perturbation schemes considering the approaches of case deletion and local influence for models with incomplete data as described in Zhu and Lee (2001) and Zhu et al. (2001).

This article is organized as follows: Sect. 2 introduces P-splines considering heavy-tailed distributions, a variant of the EM algorithm to obtain the estimators of penalized maximum likelihood (PML) is developed and presents the optimal selection of the smoothing parameter using a weighted version of the generalized cross-validation criterion (GCV). Section 3 describes the main results associated with the influence diagnostics by case deletion and local influence for models with incomplete data and presents the generalized Cook distance and the normal curvature under several perturbation schemes of the proposed model. The methodology is applied in Sect. 4 to the dataset of life expectancy in 101 countries (Leinhardt and Wasserman 1979) assuming distributions with heavier tails than the normal, also some simulation results are discussed. The numerical experiments show the utility of the proposed methodology. In Sect. 5 we present some final considerations.

2 P-splines under heavy-tailed distributions

In this section, we propose an alternative to accommodating extreme and outlying observations in penalized splines based on distributions with heavier tails than the normal. We also describe the PML estimation for P-splines using a penalized EM algorithm and present the selection of the smoothing parameter through a weighted version of the GCV criterion. The estimation of the shape parameters of the mixture variable is also described.

2.1 Estimation in P-splines using the penalized EM algorithm

Consider the model,

$$Y_i = g(x_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where the responses Y_i are observed at design points x_i and g is a smooth function defined in $[a, b]$. It is assumed that the design points are such that $a \leq x_1 < \cdots < x_n \leq b$ and the $\{\epsilon_i\}$ are random variables with zero position and scale $\phi > 0$. For simplicity, we consider that $g(x) = \sum_{j=1}^{p} a_j B_j(x)$, where p is the number of known basis functions $B_1(x), \ldots, B_p(x)$. B-splines are a common choice for the basis functions.

The class of SMN distributions (Andrews and Mallows 1974) represents an interesting alternative to the normal distribution in the presence of extreme observations and has been applied very successfully for statistical modeling by a number of authors, among them Butler et al. (1990), Lange and Sinsheimer (1993) and Jamshidian (1999). A random variable Y is said to follow an SMN distribution (Andrews and Mallows 1974) with position parameter $\mu \in \mathbb{R}$ and scale $\phi > 0$ if it can be written as $Y \overset{\mathrm{d}}{=} \mu + \tau^{-1/2} Z$, where $Z \sim \mathcal{N}(0, \phi)$ and τ is a positive random variable with distribution function $\mathcal{H}(\tau; \boldsymbol{\nu})$ where $\boldsymbol{\nu}$ represents a scalar- or vector-valued parameter that controls the shape of the distribution. The density function of Y is given by

$$f(y) = (2\pi\phi)^{-1/2} \int_0^\infty \tau^{1/2} \exp(-\tfrac{1}{2}\tau D^2) \mathrm{d}\mathcal{H}(\tau), \tag{2}$$

where $D^2 = (y-\mu)^2/\phi$ represents the distance between y to the center μ scaled by ϕ. When Y has a density given by (2) we will denote $Y \sim \mathcal{SMN}(\mu, \phi; \mathcal{H})$. It is convenient to write the distribution of the random variable Y alternatively using the following hierarchical representation:

$$Y|\tau \sim \mathcal{N}(\mu, \phi/\tau), \qquad \tau \sim \mathcal{H}(\boldsymbol{\nu}). \tag{3}$$

The formulation given in (3) is useful, for example for random number generation and parameters estimation using missing data formulation through the EM algorithm (Dempster et al. 1977). In this work the Student t and slash distributions are considered to illustrate the proposed methodology. In fact, the Student t distribution can be written using the representation in (3) considering that $\tau \sim Gamma(\nu/2, \nu/2)$ and we write $Y \sim t(\mu, \phi; \nu)$, $\nu > 0$, while that for the slash distribution, denoted by $Y \sim Slash(\mu, \phi; \nu)$, $\nu > 0$, we have $\tau \sim Beta(\nu, 1)$, $\nu > 0$. For both, the Student t and slash distributions, ν represents the degrees of freedom and this parameter control the kurtosis of the distribution. It is interesting to note that when $\nu \to \infty$ the normal distribution is recovered. It should be emphasized that other distributions also can be considered, such as the contaminated normal (Little 1988) and the Laplace or double exponential (Phillips 2002).

We will introduce scale mixtures of normal distributions for the model given in (1), by considering the following distributional assumption

$$Y_i \overset{\text{ind}}{\sim} \mathcal{SMN}(\boldsymbol{b}_i^\top \boldsymbol{a}, \phi; \mathcal{H}), \qquad i = 1, \ldots, n, \tag{4}$$

where $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)^\top = (B_j(x_i))$ is a $n \times p$ matrix and $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$. Thus, P-splines considering heavy-tailed distributions can be introduced by obtaining the PML estimates in the following penalized problem,

$$\begin{aligned} \ell_\lambda(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{obs}}) &= \ell_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{obs}}) - \frac{\lambda}{2\phi} \int_a^b \{g''(x)\}^2 \mathrm{dx} \\ &= \ell_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{obs}}) - \frac{\lambda}{2\phi} \boldsymbol{a}^\top \boldsymbol{K}^\top \boldsymbol{K} \boldsymbol{a}, \end{aligned} \tag{5}$$

where $\boldsymbol{\theta} = (\boldsymbol{a}^\top, \phi)^\top, \lambda > 0$ represents a smoothing parameter and $\boldsymbol{K}^\top \boldsymbol{K}$ is the matrix representation of the penalty described in Eilers and Marx (1996) and $\boldsymbol{K}$ is the $k \times p$ matrix ($k \leq p$) of the kth order of differencing. Details about the construction of the difference matrix $\boldsymbol{K}$, are discussed in Eilers and Marx (1996, 2010). The log-likelihood function for the class of SMN distributions given in (5) assumes the form

$$\ell_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{obs}}) = -\frac{n}{2}\log 2\pi\phi - \sum_{i=1}^{n} \log \int_0^\infty \tau_i^{1/2} \exp\big(-\tfrac{1}{2}\tau_i D_i^2(\boldsymbol{\theta})\big) \mathrm{d}\mathcal{H}(\tau_i),$$

where $D_i^2(\boldsymbol{\theta}) = (Y_i - \boldsymbol{b}_i^\top \boldsymbol{a})^2/\phi$, for $i = 1, \ldots, n$.

The estimation problem given in (5) can be significantly simplified by considering an incomplete data formulation. Using the hierarchical formulation of a random variable with an SMN distribution, it is possible to re-write the model proposed in (4) as

$$Y_i|\tau_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{b}_i^\top \boldsymbol{a}, \phi/\tau_i), \qquad \tau_i \overset{\text{ind}}{\sim} \mathcal{H}(\boldsymbol{\nu}), \qquad i = 1, \ldots, n.$$

Thus, it is possible to apply the penalized EM algorithm (Green 1990) to estimate the parameters in (5) by assuming that $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^\top$ are missing variables. The penalized log-likelihood function for the model based on complete data $\boldsymbol{Y}_{\mathsf{com}} = (\boldsymbol{Y}^\top, \boldsymbol{\tau}^\top)^\top$ is defined through

$$\ell_\lambda(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{com}}) = \ell_{\mathsf{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{com}}) - \frac{\lambda}{2\phi}\boldsymbol{a}^\top \boldsymbol{K}^\top \boldsymbol{K}\boldsymbol{a},$$

with

$$\begin{aligned}\ell_{\mathsf{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{com}}) &= -\frac{n}{2}\log\phi - \frac{1}{2\phi}\sum_{i=1}^{n}\tau_i(Y_i - \boldsymbol{b}_i^\top \boldsymbol{a})^2 + \log h^{(n)}(\boldsymbol{\tau}; \boldsymbol{\nu}) \\ &= -\frac{n}{2}\log\phi - \frac{1}{2\phi}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{a})^\top \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{a}) + \log h^{(n)}(\boldsymbol{\tau}; \boldsymbol{\nu}),\end{aligned}$$

where $h^{(n)}(\boldsymbol{\tau}; \boldsymbol{\nu})$ is the joint density function of the mixture variables $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^\top$ and $\boldsymbol{W} = \mathrm{diag}(\tau_1, \ldots, \tau_n)$. Assuming that $\boldsymbol{\nu}$ is known, it is possible to show that the conditional expectation of the complete-data-penalized log-likelihood function considering a current estimate $\boldsymbol{\theta}^{(k)}$, given by

$$Q_\lambda(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \mathrm{E}\left[\ell_{\mathsf{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{com}})|\boldsymbol{Y}, \boldsymbol{\theta}^{(k)}\right] - \frac{\lambda}{2\phi}\boldsymbol{a}^\top \boldsymbol{K}^\top \boldsymbol{K}\boldsymbol{a},$$

can be expressed as

$$Q_\lambda(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = -\frac{n}{2}\log\phi - \frac{1}{2\phi}\left[(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{a})^\top \boldsymbol{W}^{(k)}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{a}) + \lambda \boldsymbol{a}^\top \boldsymbol{K}^\top \boldsymbol{K}\boldsymbol{a}\right], \quad (6)$$

where $\boldsymbol{W}^{(k)} = \text{diag}(\tau_1^{(k)}, \ldots, \tau_n^{(k)})$ with $\tau_i^{(k)} = \text{E}(\tau_i | Y_i, \boldsymbol{\theta}^{(k)})$. In general, it is possible to show that the weight function, defined by the expectation $\tau_i^{(k)}$ is given by

$$\text{E}(\tau_i | Y_i, \boldsymbol{\theta}^{(k)}) = \left. \frac{\int_0^\infty \tau_i^{3/2} \exp(-\frac{1}{2}\tau_i D_i^2(\boldsymbol{\theta}))\text{d}\mathcal{H}(\tau_i)}{\int_0^\infty \tau_i^{1/2} \exp(-\frac{1}{2}\tau_i D_i^2(\boldsymbol{\theta}))\text{d}\mathcal{H}(\tau_i)} \right|_{\theta=\theta^{(k)}}.$$

Note that for most of the distributions in the SMN class, the weight function $\tau_i^{(k)}$ can be easily computed (see Lange and Sinsheimer (1993)). For the Student t distribution is well known that

$$\tau_i^{(k)} = (\nu + 1)/(\nu + D_i^2(\boldsymbol{\theta}^{(k)})), \tag{7}$$

while for the slash distribution, $\tau_i^{(k)}$ assumes the form (Jamshidian 1999)

$$\tau_i^{(k)} = \left(\frac{2\nu + 1}{D_i^2(\boldsymbol{\theta}^{(k)})} \right) \frac{P(\nu + \frac{3}{2}, D_i^2(\boldsymbol{\theta}^{(k)})/2)}{P(\nu + \frac{1}{2}, D_i^2(\boldsymbol{\theta}^{(k)})/2)}, \tag{8}$$

where $P(\alpha, z)$ is the incomplete gamma function of parameter α at z (Abramowitz and Stegun 1970, p. 260), defined as

$$P(\alpha, z) = \frac{1}{\Gamma(\alpha)} \int_0^z e^{-t} t^{\alpha-1} \text{dt}.$$

To maximize $Q_\lambda(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ given in (6) with respect to $\boldsymbol{\theta} = (\boldsymbol{a}^\top, \phi)^\top$, we solve the first-order condition and update $\boldsymbol{\theta}^{(k+1)}$ as

$$\boldsymbol{a}_\lambda^{(k+1)} = (\boldsymbol{B}^\top \boldsymbol{W}^{(k)} \boldsymbol{B} + \lambda \boldsymbol{K}^\top \boldsymbol{K})^{-1} \boldsymbol{B}^\top \boldsymbol{W}^{(k)} \boldsymbol{Y}, \tag{9}$$

$$\phi_\lambda^{(k+1)} = \frac{1}{n} \left\{ RSS_{W^{(k)}}(\boldsymbol{a}_\lambda^{(k+1)}) + \lambda \| \boldsymbol{K} \boldsymbol{a}_\lambda^{(k+1)} \|^2 \right\}, \tag{10}$$

where $RSS_W(\boldsymbol{a}) = (\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{a})^\top \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{a})$. The PML estimates for the problem in (5) are obtained by iterating the E and M steps of the algorithm, described in Eqs. (6), (9) and (10) until reaching convergence.

It is possible to modify the estimation procedure delineated above to simultaneously estimate $\boldsymbol{a}$, ϕ and the tuning parameter $\boldsymbol{\nu}$. In this case, the expectation of the complete-data-penalized log-likelihood function assumes the form

$$Q_\lambda(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = Q_\lambda(\boldsymbol{a}, \phi|\boldsymbol{\theta}^{(k)}) + Q_\lambda(\boldsymbol{\nu}|\boldsymbol{\theta}^{(k)}), \tag{11}$$

with $\boldsymbol{\theta} = (\boldsymbol{a}^\top, \phi, \boldsymbol{\nu}^\top)^\top$, where $Q_\lambda(\boldsymbol{a}, \phi|\boldsymbol{\theta}^{(k)})$ is given in equation (6), while

$$Q_\lambda(\boldsymbol{\nu}|\boldsymbol{\theta}^{(k)}) = \text{E}[\log h^{(n)}(\boldsymbol{\tau}; \boldsymbol{\nu})|\boldsymbol{Y}, \boldsymbol{\theta}^{(k)}]. \tag{12}$$

For the M step we update $\boldsymbol{a}^{(k+1)}$ and $\phi^{(k+1)}$ according to Eqs. (9) and (10), respectively, and we obtain $\boldsymbol{\nu}^{(k+1)}$ through

$$\boldsymbol{\nu}^{(k+1)} = \arg\max_{\nu} Q_\lambda(\boldsymbol{\nu}|\boldsymbol{\theta}^{(k)}). \tag{13}$$

Indeed $\boldsymbol{\nu}^{(k+1)}$ can be obtained as the solution of the system of equations

$$\partial Q_\lambda(\boldsymbol{\nu}|\boldsymbol{\theta}^{(k)})/\partial \boldsymbol{\nu} = \mathbf{0}.$$

Hence, the form that conditional expectation (12) and the M step (13) adopt depends on the specific choice of the distribution of the random variable τ_i.

To illustrate the estimation of the tuning parameter $\boldsymbol{\nu}$, below we present the estimation of the degrees of freedom for the slash and Student t distributions. Additional details about this respect can be found in Lange and Sinsheimer (1993), McLachlan and Krishnan (1997) and Jamshidian (1999).

For the particular case of the Student t distribution, it follows that $\tau_i|Y_i \overset{\text{ind}}{\sim} Gamma((\nu+1)/2, (\nu + D_i^2(\boldsymbol{\theta}))/2)$, for $i = 1, \ldots, n$. Using results from McLachlan and Krishnan (1997) we have

$$\mathrm{E}(\log \tau_i|Y_i, \boldsymbol{\theta}^{(k)}) = \log \tau_i^{(k)} + \left\{\psi\left(\frac{\nu^{(k)}+1}{2}\right) - \log\left(\frac{\nu^{(k)}+1}{2}\right)\right\},$$

where $\tau_i^{(k)}$ is defined in (7) and $\psi(z) = \mathrm{d}\log\Gamma(z)/\mathrm{d}z$ is the digamma function (Abramowitz and Stegun 1970, p. 268). Thus, the conditional expectation of the complete-data-penalized log-likelihood associated with ν, assumes the form

$$\begin{aligned} Q_\lambda(\nu|\boldsymbol{\theta}^{(k)}) = \frac{n\nu}{2}\log\left(\frac{\nu}{2}\right) - n\log\Gamma\left(\frac{\nu}{2}\right) + \frac{n\nu}{2}\Big\{\frac{1}{n}\sum_{i=1}^{n}(\log(\tau_i^{(k)}) - \tau_i^{(k)}) \\ + \psi\left(\frac{\nu^{(k)}+1}{2}\right) - \log\left(\frac{\nu^{(k)}+1}{2}\right)\Big\}. \end{aligned}$$

It is possible to update $\nu^{(k+1)}$ as the solution to equation $\partial Q_\lambda(\nu|\boldsymbol{\theta}^{(k)})/\partial\nu = 0$ using an one-dimensional Newton–Raphson method.

Using errors following a slash distribution, the calculation of the conditional expectation in (12), requires evaluating (see Lange and Sinsheimer 1993)

$$\begin{aligned} \mathrm{E}(\log \tau_i|Y_i, \boldsymbol{\theta}^{(k)}) &= \frac{\int_0^1 \log(\tau_i)\tau_i^{\nu-1/2}\exp(-\frac{1}{2}\tau_i D_i^2(\boldsymbol{\theta}^{(k)}))\mathrm{d}\tau_i}{\int_0^1 \tau_i^{\nu-1/2}\exp(-\frac{1}{2}\tau_i D_i^2(\boldsymbol{\theta}^{(k)}))\mathrm{d}\tau_i} \\ &= \psi(\nu+\tfrac{1}{2}) - \log(D_i^2(\boldsymbol{\theta}^{(k)})/2) + \frac{\partial P(\nu+\frac{1}{2}, D_i^2(\boldsymbol{\theta}^{(k)})/2)/\partial\nu}{P(\nu+\frac{1}{2}, D_i^2(\boldsymbol{\theta}^{(k)})/2)}. \end{aligned}$$

The derivative of the incomplete gamma function $\partial P(a, x)/\partial a$ can be evaluated using the algorithm described in Moore (1982). In this case, the conditional expectation in (12) is given by

$$Q_\lambda(\nu|\boldsymbol{\theta}^{(k)}) = n\log\nu + \nu\sum_{i=1}^{n}\mathrm{E}(\log\tau_i|Y_i,\boldsymbol{\theta}^{(k)}).$$

Maximizing $Q_\lambda(\nu|\boldsymbol{\theta}^{(k)})$ in relation to ν, we obtain

$$\nu^{(k+1)} = -\frac{n}{\sum_{i=1}^{n}\mathrm{E}(\log\tau_i|Y_i,\boldsymbol{\theta}^{(k)})}.$$

Remark 1 In this work, we address the estimation of the shape parameters for the mixture variables using the EM algorithm following the approach proposed for a number of authors in settings like nonlinear regression models (Lange and Sinsheimer 1993; Jamshidian 1999) and linear mixed-effects models under Student t errors (Pinheiro et al. 2001; Lin and Lee 2006). Although the approach of these works has been quite successful in practice, some authors (see, for instance Lucas 1997; Fernández and Steel 1999) have warned about potential problems that may arise in the estimation of degrees of freedom for the Student t distribution. Particularly, Lucas (1997) has pointed out using influence functions for the univariate case that the protection against outliers is only attained when this parameter is kept fixed. Moreover, when the degrees of freedom is estimated by maximum likelihood the influence function for the scale, degrees of freedom and the change-of-variance of the position parameter is unbounded. Thus, one alternative is to assume that the parameters associated with the mixture variables τ_i are known. To achieve protection against outliers Lange et al. (1989) suggest that the degrees of freedom of the Student t distribution must be kept fixed in a small reasonable value such as $\nu = 4$. There is an option in the heavy package that allows one to keep the shape parameter $\boldsymbol{\nu}$ fixed.

2.2 Smoothing parameter selection

Several authors have suggested modifications to the GCV criterion (Craven and Wahba 1979) for the appropriate selection of the smoothing parameter λ. For example, O'Sullivan et al. (1986) and Gu (1992) proposed versions of the GCV criterion for non-Gaussian data focused mainly on the penalized maximum likelihood for distributions in the exponential family, while Wei (2005) examined the asymptotic properties of the criterion of robust cross validation based on M-estimation procedures. In this work, we choose the smoothing parameter minimizing the weighted cross-validation criterion as defined by O'Sullivan et al. (1986) as

$$V(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n}\widehat{\tau}_i(Y_i - \widehat{g}_\lambda(x_i))^2}{\{\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}_{\widehat{W}}(\lambda))/n\}^2} = \frac{\|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{I} - \boldsymbol{H}_{\widehat{W}}(\lambda))\boldsymbol{Y}\|^2/n}{\{\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}_{\widehat{W}}(\lambda))/n\}^2}, \tag{14}$$

with $\widehat{\tau}_i = \mathrm{E}(\tau_i|Y_i,\widehat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{g}}_\lambda = (\widehat{g}_\lambda(x_1),\ldots,\widehat{g}_\lambda(x_n))^\top = \boldsymbol{H}_{\widehat{W}}(\lambda)\boldsymbol{Y}$, where the prediction matrix assumes the form

$$\boldsymbol{H}_W(\lambda) = \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{W}\boldsymbol{B} + \lambda\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{B}^\top\boldsymbol{W}. \tag{15}$$

As Gu (1992) suggested, it is possible to alternate the minimization of (14) with the steps of the penalized EM algorithm described in equations (6) and (9)–(10) or (9)–(13), whichever applies.

Remark 2 The convergence properties of the penalized EM algorithm proposed in Sect. 2.1 have been studied in the general context of penalized log-likelihood estimation for λ fixed by Green (1990). However, for varying λ Gu (1992), Xiang and Wahba (1996) and Gu and Xiang (2001) among others, have discussed that choosing the smoothing parameter via the indirect method described above may not converge. In our implementation, we follow the suggestions given by Gu (1992). Thus, we did not find such problems in our numerical experiments. In addition, can be stressed that the WGCV criterion defined in (14) is similar to the robust GCV used by (Tharmaratnam et al., 2010, Eq. 2.18). In fact, following Lucas (1997) the relationship between the criteria WGCV and robust GCV can be highlighted by defining $\widehat{\boldsymbol{a}}_\lambda$ as the solution to

$$\min_{a} \sum_{i=1}^{n} \rho(D_i^2) + \frac{\lambda}{2\phi} \boldsymbol{a}^\top \boldsymbol{K}^\top \boldsymbol{K} \boldsymbol{a},$$

where $\rho(D^2) = -\log f(y; \boldsymbol{a}, \phi)$, with $f(y; \boldsymbol{a}, \phi)$ the density function obtained from Eq. (4). Furthermore, we can also note that the matrix $\boldsymbol{H}_{\widehat{W}}(\lambda)$ have the same diagonal elements than the following matrix

$$\widehat{\boldsymbol{W}}^{1/2} \boldsymbol{B} (\boldsymbol{B}^\top \widehat{\boldsymbol{W}} \boldsymbol{B} + \lambda \boldsymbol{K}^\top \boldsymbol{K})^{-1} \boldsymbol{B}^\top \widehat{\boldsymbol{W}}^{1/2},$$

which is analogous to the prediction matrix defined by Tharmaratnam et al. (2010). The implementation available in the heavy package uses two nested singular value decompositions to efficiently evaluate the weighted GCV criterion, details are presented in the Appendix A of the supplementary material.

3 Influence diagnostics

Below we describe two of the main procedures to determine the influence of outlying observations. We consider diagnostic measures suitable for models with incomplete data, based on the PML estimation using the penalized EM algorithm. First, we present the approach of case deletion using the generalized Cook distance (Zhu et al. 2001). Subsequently, we develop the diagnostic using the local influence method proposed by Zhu and Lee (2001).

The proofs of Propositions 1–5 are deferred to Appendix B of the supplementary material.

3.1 Case deletion measures

To evaluate the effect of dropping the ith observation on the PML estimation of the p^*-dimensional parameter vector $\boldsymbol{\theta}$, it is possible to use the Cook distance, defined as

$$C_i = (\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})^\top \boldsymbol{M} (\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}), \qquad i = 1, \ldots, n,$$

where $\widehat{\boldsymbol{\theta}}_{(i)}$ represents the PML estimate of $\boldsymbol{\theta}$ once the ith observation has been dropped from the dataset and $\boldsymbol{M}$ is a positive definite matrix of order $p^* \times p^*$ (see Cook and Weisberg 1982). Zhu et al. (2001) proposed the generalized Cook distance for models with incomplete data as defined by

$$GC_i = (\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})^\top \{-\ddot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}), \qquad i = 1, \ldots, n,$$

where $\ddot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \partial^2 Q_\lambda(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top|_{\theta=\widehat{\theta}}$. To reduce the computational burden involved in calculating $\widehat{\boldsymbol{\theta}}_{(i)}$, $i = 1, \ldots, n$, the following one-step approximation has been proposed (Cook and Weisberg 1982; Zhu et al. 2001)

$$\widehat{\boldsymbol{\theta}}^1_{(i)} = \widehat{\boldsymbol{\theta}} + \{-\ddot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\dot{Q}_{\lambda(i)}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}), \qquad i = 1, \ldots, n,$$

where

$$Q_{\lambda(i)}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \mathrm{E}[\ell_{\mathsf{C}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{com}(i)})|\boldsymbol{Y}_{(i)}, \widehat{\boldsymbol{\theta}}] - \lambda J(\boldsymbol{\theta}),$$

with $\boldsymbol{Y}_{\mathsf{com}(i)} = (\boldsymbol{Y}^\top_{(i)}, \boldsymbol{\tau}^\top_{(i)})^\top$ being the complete-data vector when the ith observation has been deleted and $\dot{Q}_{\lambda(i)}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \partial Q_{\lambda(i)}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}|_{\theta=\widehat{\theta}}$. For a wide variety of statistical models it is possible to write $Q_\lambda(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^n Q_{\lambda,i}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$. Thus, $Q_{\lambda(i)}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \sum_{j\neq i} Q_{\lambda,j}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$, in which case we have

$$\dot{Q}_{\lambda(i)}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) + \dot{Q}_{\lambda,i}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \dot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \mathbf{0},$$

and consequently we consider the following one-step approximation for the generalized Cook distance

$$GC^1_i = \dot{Q}^\top_{\lambda,i}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\{-\ddot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\dot{Q}_{\lambda,i}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}), \qquad i = 1, \ldots, n.$$

The following proposition gives the analytical form of the Hessian matrix for the penalized splines under heavy-tailed distributions discussed in Sect. 2.

Proposition 1 *For the model given in Eq. (4) from Sect. 2.1 the* $(p+1) \times (p+1)$ *Hessian matrix associated with the penalized* $Q_\lambda(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ *function evaluated at* $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ *assumes the form*

$$\ddot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} \ddot{Q}_\lambda(\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{\theta}}) & \mathbf{0} \\ \mathbf{0} & \ddot{Q}_\lambda(\widehat{\phi}|\widehat{\boldsymbol{\theta}}) \end{pmatrix} = -\frac{1}{\widehat{\phi}} \begin{pmatrix} \boldsymbol{B}^\top \widehat{\boldsymbol{W}} \boldsymbol{B} + \lambda \boldsymbol{K}^\top \boldsymbol{K} & \mathbf{0} \\ \mathbf{0} & n/(2\widehat{\phi}) \end{pmatrix}.$$

To obtain the generalized Cook distance in the model described in Sect. 2.1, we consider $Q_\lambda(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^n Q_{\lambda,i}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$, with

$$Q_{\lambda,i}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = -\frac{1}{2}\log\phi + \frac{1}{2\phi}\left\{\widehat{\tau}_i(Y_i - \boldsymbol{b}_i^\top \boldsymbol{a})^2 + \frac{\lambda}{n}\|\boldsymbol{K}\boldsymbol{a}\|^2\right\}.$$

Using the Proposition 1 follows immediately that the one-step approximation for the generalized Cook distance is given as $GC_i^1 = GC_i^1(\widehat{\boldsymbol{a}}) + GC_i^1(\widehat{\phi})$, $i = 1, \ldots, n$, where

$$\begin{aligned} GC_i^1(\widehat{\boldsymbol{a}}) &= \dot{Q}_{\lambda,i}^{\top}(\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{\theta}})\{-\ddot{Q}_{\lambda}(\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{\theta}})\}^{-1}\dot{Q}_{\lambda,i}(\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{\theta}}), \\ GC_i^1(\widehat{\phi}) &= \dot{Q}_{\lambda,i}^{\top}(\widehat{\phi}|\widehat{\boldsymbol{\theta}})\{-\ddot{Q}_{\lambda}(\widehat{\phi}|\widehat{\boldsymbol{\theta}})\}^{-1}\dot{Q}_{\lambda,i}(\widehat{\phi}|\widehat{\boldsymbol{\theta}}), \end{aligned} \tag{16}$$

with

$$\begin{aligned} \dot{Q}_{\lambda,i}(\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{\theta}}) &= -\frac{1}{\widehat{\phi}}\left\{\widehat{\tau}_i(Y_i - \boldsymbol{b}_i^{\top}\widehat{\boldsymbol{a}})\boldsymbol{b}_i + \frac{\lambda\widehat{\phi}}{n}\boldsymbol{K}^{\top}\boldsymbol{K}\widehat{\boldsymbol{a}}\right\}, \\ \dot{Q}_{\lambda,i}(\widehat{\phi}|\widehat{\boldsymbol{\theta}}) &= -\frac{1}{2\widehat{\phi}} - \frac{1}{2\widehat{\phi}^2}\left\{\widehat{\tau}_i(Y_i - \boldsymbol{b}_i^{\top}\widehat{\boldsymbol{a}})^2 + \frac{\lambda}{n}\|\boldsymbol{K}\widehat{\boldsymbol{a}}\|^2\right\}. \end{aligned}$$

The distances $GC_i^1(\widehat{\boldsymbol{a}})$ and $GC_i^1(\widehat{\phi})$ given in (16) offer an interesting interpretation. In fact, $GC_i^1(\widehat{\boldsymbol{a}})$ allows to assess the influence of the ith observation on the PML estimate of $\boldsymbol{a}$ and analogously for $GC_i^1(\widehat{\phi})$. In addition, these measures complement and extend the results developed by Eubank and Gunst (1986); Kim (1996); Wei (2004) and Ibacache-Pulgar and Paula (2011).

3.2 Local influence

The local influence method proposed by Cook (1986) allows studying the effect produced by introducing small perturbations on the model and/or the data. The procedure was extended by Zhu and Lee (2001) to manipulate situations with incomplete data. They were focused on assessing the local behavior of the Q-displacement function given by

$$f_Q(\boldsymbol{\omega}) = 2\{Q_{\lambda}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) - Q_{\lambda}(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})|\widehat{\boldsymbol{\theta}})\},$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_q)^{\top}$ is a vector of perturbations restricted to some open subset $\Omega \subset \mathbb{R}^q$ and in the context of this work, $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$ denotes the PML estimate of $\boldsymbol{\theta}$ based on

$$Q_{\lambda}(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \mathrm{E}[\ell_{\mathsf{c}}(\boldsymbol{\theta}, \boldsymbol{\omega}; \boldsymbol{Y}_{\mathsf{com}})|\boldsymbol{Y}_{\mathsf{obs}}, \widehat{\boldsymbol{\theta}}] - \lambda J(\boldsymbol{\theta}).$$

It is assumed that there is a vector of null perturbation $\boldsymbol{\omega}_0 \in \Omega$ which satisfies $\ell_{\mathsf{o}}(\boldsymbol{\theta}, \boldsymbol{\omega}_0; \boldsymbol{Y}_{\mathsf{obs}}) = \ell_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{obs}})$ and $\ell_{\mathsf{c}}(\boldsymbol{\theta}, \boldsymbol{\omega}_0; \boldsymbol{Y}_{\mathsf{com}}) = \ell_{\mathsf{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{com}})$.

The objective of the local influence technique is to compare $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$ by studying the local behavior of $\boldsymbol{\gamma}(\boldsymbol{\omega}) = (\boldsymbol{\omega}^{\top}, f_Q(\boldsymbol{\omega}))^{\top}$ around $\boldsymbol{\omega}_0$ (Cook 1986; Zhu and Lee 2001). Consider $\boldsymbol{\omega} = \boldsymbol{\omega}_0 + \varepsilon\boldsymbol{h}$, where $\boldsymbol{h}$, $\|\boldsymbol{h}\| = 1$ is an unitary direction and $\varepsilon \in \mathbb{R}$. Zhu and Lee (2001) used the same reasoning developed by Cook (1986) and showed that the normal curvature $C_h(\boldsymbol{\theta})$ can be employed to characterize the local behavior of $f_Q(\boldsymbol{\omega}_0 + \varepsilon\boldsymbol{h})$ around the value $\varepsilon = 0$ for a direction $\boldsymbol{h}$, given by

$$C_h(\boldsymbol{\theta}) = 2\,\boldsymbol{h}^\top \boldsymbol{\Delta}^\top(\boldsymbol{\omega}_0)\{-\ddot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\boldsymbol{\Delta}(\boldsymbol{\omega}_0)\boldsymbol{h},$$

where $\boldsymbol{\Delta}(\boldsymbol{\omega}) = \partial^2 Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^\top|_{\theta=\widehat{\theta}(\omega)}$.

The direction of maximum curvature $\boldsymbol{h}_{\max}$, determined by the vector associated with the largest eigenvalue of the matrix $\boldsymbol{F} = \boldsymbol{\Delta}^\top(\boldsymbol{\omega}_0)\{-\ddot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\boldsymbol{\Delta}(\boldsymbol{\omega}_0)$ is used to identify how to perturb the postulated model to obtain the greatest local change in the Q-displacement function.

Several authors have proposed examining other relevant directions to investigate local influence. For example Escobar and Meeker (1992) suggested considering the index plot of $C_i(\boldsymbol{\theta}) = C_{h_i}(\boldsymbol{\theta}), i = 1, \ldots, n$, where $\boldsymbol{h}_i$ is a $q \times 1$ vector with one in the ith position and zeros elsewhere. In fact, $C_i(\boldsymbol{\theta})$ allows the evaluation of the influence of the ith observation due to the aggregated contribution of all the basic perturbation vectors (Poon and Poon 1999). Poon and Poon (1999) proposed the conformal normal curvature $B_h(\boldsymbol{\theta}) = C_h(\boldsymbol{\theta})/\{\mathrm{tr}(\boldsymbol{T}^2)\}^{1/2}$ to avoid invariance problems under uniform changes of scale. The conformal normal curvature satisfies that $0 \leq B_h(\boldsymbol{\theta}) \leq 1$, a property that allows comparison of curvatures obtained by considering different SMN models.

Following Thomas (1991) and Shi and Wang (1999) it is possible to determine the observations that have a strong impact on the smoothing parameter selection in penalized splines, that is $\widehat{\lambda}(\boldsymbol{\omega})$, obtained by introducing a small perturbation $\boldsymbol{\omega} \in \Omega$ on the WGCV criterion given in (14). It is assumed that there is an $\boldsymbol{\omega}_0 \in \Omega$ vector of no perturbation, such that $\widehat{\lambda}(\boldsymbol{\omega}_0) = \widehat{\lambda}$. The direction of the greatest local change is $\boldsymbol{h}_{\max}(V) \propto \partial\widehat{\lambda}(\boldsymbol{\omega})/\partial\boldsymbol{\omega}$, which should be evaluated at $\boldsymbol{\omega}_0$. Since $\widehat{\lambda}(\boldsymbol{\omega})$ is chosen minimizing a perturbed version of the WGCV criterion, we have $\partial V(\lambda, \boldsymbol{\omega})/\partial\lambda|_{\lambda=\widehat{\lambda}(\omega)} = 0$. Differentiating both sides of this equation with respect to $\boldsymbol{\omega}$ and evaluating this derivative at $\boldsymbol{\omega}_0$, gives

$$\left\{\frac{\partial^2 V(\lambda, \boldsymbol{\omega})}{\partial\boldsymbol{\omega}\partial\lambda} + \frac{\partial^2 V(\lambda, \boldsymbol{\omega})}{\partial\lambda^2}\frac{\partial\widehat{\lambda}(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}}\right\}\Big|_{\omega=\omega_0, \lambda=\widehat{\lambda}} = 0.$$

Thus,

$$\frac{\partial\widehat{\lambda}(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}}\Big|_{\omega=\omega_0} = \left\{-\left(\frac{\partial^2 V(\lambda, \boldsymbol{\omega})}{\partial\lambda^2}\right)^{-1}\frac{\partial^2 V(\lambda, \boldsymbol{\omega})}{\partial\boldsymbol{\omega}\partial\lambda}\right\}\Big|_{\omega=\omega_0, \lambda=\widehat{\lambda}}. \tag{17}$$

Below we consider the scale and response perturbation schemes. Each scheme was applied on the complete-data-penalized log-likelihood function and the WGCV criterion. To find $\boldsymbol{h}_{\max}$ we must compute the curvature matrix $\boldsymbol{F} = \boldsymbol{\Delta}^\top(\boldsymbol{\omega}_0)\{-\ddot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}$ $\boldsymbol{\Delta}(\boldsymbol{\omega}_0)$, where $\ddot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$ was given in Proposition 1 and for each perturbation scheme $\boldsymbol{\Delta}(\boldsymbol{\omega})$ assumes the partitioned form $\boldsymbol{\Delta}(\boldsymbol{\omega}) = (\boldsymbol{\Delta}_a^\top(\boldsymbol{\omega}), \boldsymbol{\Delta}_\phi^\top(\boldsymbol{\omega}))^\top$, with

$$\boldsymbol{\Delta}_a(\boldsymbol{\omega}) = \frac{\partial^2 Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{a}\partial\boldsymbol{\omega}^\top}, \qquad \boldsymbol{\Delta}_\phi(\boldsymbol{\omega}) = \frac{\partial^2 Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial\phi\partial\boldsymbol{\omega}^\top}.$$

We should emphasize that the $\boldsymbol{\Delta}(\boldsymbol{\omega})$ matrix is specific for the perturbation scheme under consideration. Propositions 2 and 3 present analytic expressions for the matrix

$\boldsymbol{\Delta}(\boldsymbol{\omega}_0)$, while Propositions 4 and 5 give explicit formulas for $\partial\widehat{\lambda}(\boldsymbol{\omega})/\partial\boldsymbol{\omega}$ when the WGCV criterion is perturbed.

3.2.1 Scale perturbation on complete-data-penalized log-likelihood

This perturbation scheme is defined by introducing weights in the scale, that is, the following distributional assumption is considered

$$Y_i(\boldsymbol{\omega}) \overset{\text{ind}}{\sim} \mathcal{SMN}(\boldsymbol{b}_i^\top \boldsymbol{a}, \phi/\omega_i; \mathcal{H}), \qquad i = 1, \ldots, n, \tag{18}$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$, with $\omega_i > 0$ for $i = 1, \ldots, n$. In this case, the vector of null perturbation is $\boldsymbol{\omega}_0 = \mathbf{1}_n$, with $\mathbf{1}_n = (1, \ldots, 1)^\top$. The perturbed log-likelihood function for the complete-data model assumes the form

$$\ell_{\mathsf{c}}(\boldsymbol{\theta}, \boldsymbol{\omega}; \boldsymbol{Y}_{\mathsf{com}}) = -\frac{n}{2}\log\phi - \frac{1}{2\phi}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{a})^\top \boldsymbol{W}^{1/2}\text{diag}(\boldsymbol{\omega})\boldsymbol{W}^{1/2}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{a}),$$

where $\text{diag}(\boldsymbol{\omega}) = \text{diag}(\omega_1, \ldots, \omega_n)$ is a diagonal matrix, whose diagonal elements are given by the vector $\boldsymbol{\omega}$. Then the following proposition holds.

Proposition 2 *For the penalized splines considering heavy-tailed distributions, under the scale perturbation scheme defined in Eq. (18) and when $\boldsymbol{\theta} = (\boldsymbol{a}^\top, \phi)^\top$ are the parameters of interest, the $\boldsymbol{\Delta}(\boldsymbol{\omega}_0)$ matrix can be written as $\boldsymbol{\Delta}(\boldsymbol{\omega}_0) = (\boldsymbol{\Delta}_a^\top(\boldsymbol{\omega}_0), \boldsymbol{\Delta}_\phi^\top(\boldsymbol{\omega}_0))^\top$, where*

$$\boldsymbol{\Delta}_a(\boldsymbol{\omega}_0) = \frac{1}{\widehat{\phi}}\boldsymbol{B}^\top\widehat{\boldsymbol{W}}\text{diag}(\boldsymbol{e}), \qquad \boldsymbol{\Delta}_\phi(\boldsymbol{\omega}_0) = \frac{1}{2\widehat{\phi}}\boldsymbol{e}^\top\widehat{\boldsymbol{W}}\text{diag}(\boldsymbol{e}),$$

with $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{a}}_\lambda$ being the residual vector.

A direct consequence of Proposition 2 is that the curvature matrix $\boldsymbol{F}$ can be written as $\boldsymbol{F} = \boldsymbol{F}_a + \boldsymbol{F}_\phi$, with

$$\begin{aligned}\boldsymbol{F}_a &= \boldsymbol{\Delta}_a^\top(\boldsymbol{\omega}_0)\{-\ddot{Q}_\lambda(\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{\theta}})\}^{-1}\boldsymbol{\Delta}_a(\boldsymbol{\omega}_0)\\ &= \frac{1}{\widehat{\phi}}\text{diag}(\boldsymbol{e})\widehat{\boldsymbol{W}}\boldsymbol{B}(\boldsymbol{B}^\top\widehat{\boldsymbol{W}}\boldsymbol{B} + \lambda\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{B}^\top\widehat{\boldsymbol{W}}\text{diag}(\boldsymbol{e}),\end{aligned}$$

and

$$\boldsymbol{F}_\phi = \boldsymbol{\Delta}_\phi^\top(\boldsymbol{\omega}_0)\{-\ddot{Q}_\lambda(\widehat{\phi}|\widehat{\boldsymbol{\theta}})\}^{-1}\boldsymbol{\Delta}_\phi(\boldsymbol{\omega}_0) = \frac{1}{2n}\text{diag}(\boldsymbol{e})\widehat{\boldsymbol{W}}\boldsymbol{e}\boldsymbol{e}^\top\widehat{\boldsymbol{W}}\text{diag}(\boldsymbol{e}).$$

This perturbation scheme is equivalent to the case-weights perturbation, where weights are introduced with the aim to detect which observations have a prominent contribution on the residual sum of squares, $RSS_W(\boldsymbol{a})$. Indeed, the case-weights (or scale) perturbation generalizes the concept of influence by means of cases deletion (see, for instance Thomas 1991).

3.2.2 Response perturbation on complete-data-penalized log-likelihood

This scheme is defined by introducing additive perturbations in the observed responses as $\boldsymbol{Y}(\boldsymbol{\omega}) = \boldsymbol{Y} + \boldsymbol{\omega}$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ and $\boldsymbol{\omega}_0 = \mathbf{0}$ denote the vector of null perturbation. Under this perturbation scheme, the conditional expectation of the complete-data-penalized log-likelihood function is given by

$$Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = -\frac{n}{2}\log\phi - \frac{1}{2\phi}[(\boldsymbol{Y}(\boldsymbol{\omega}) - \boldsymbol{B}\boldsymbol{a})^\top \widehat{\boldsymbol{W}}(\boldsymbol{Y}(\boldsymbol{\omega}) - \boldsymbol{B}\boldsymbol{a}) + \lambda \boldsymbol{a}^\top \boldsymbol{K}^\top \boldsymbol{K}\boldsymbol{a}].$$

To obtain an explicit formulae of the $\boldsymbol{\Delta}(\boldsymbol{\omega}_0)$ matrix under response variable perturbation consider the following proposition.

Proposition 3 *For penalized splines assuming the class of scale mixture of normal distributions, under the response perturbation scheme and considering that $\boldsymbol{\theta} = (\boldsymbol{a}^\top, \phi)^\top$ are the parameters of interest, then the $\boldsymbol{\Delta}(\boldsymbol{\omega}_0)$ matrix can be expressed as $\boldsymbol{\Delta}(\boldsymbol{\omega}_0) = (\boldsymbol{\Delta}_a^\top(\boldsymbol{\omega}_0), \boldsymbol{\Delta}_\phi^\top(\boldsymbol{\omega}_0))^\top$, where*

$$\boldsymbol{\Delta}_a(\boldsymbol{\omega}_0) = \frac{1}{\widehat{\phi}}\boldsymbol{B}^\top\widehat{\boldsymbol{W}}, \qquad \boldsymbol{\Delta}_\phi(\boldsymbol{\omega}_0) = \frac{1}{\widehat{\phi}}\boldsymbol{e}^\top\widehat{\boldsymbol{W}},$$

and $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{a}}_\lambda$ is the residual vector.

Note that when ϕ is known, the curvature matrix for the response perturbation scheme assumes the form

$$\begin{aligned}\boldsymbol{F} &= \boldsymbol{\Delta}_a^\top(\boldsymbol{\omega}_0)\{-\ddot{Q}_\lambda(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\boldsymbol{\Delta}_a(\boldsymbol{\omega}_0)\\ &= \frac{1}{\phi}\widehat{\boldsymbol{W}}\boldsymbol{B}(\boldsymbol{B}^\top\widehat{\boldsymbol{W}}\boldsymbol{B} + \lambda\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{B}^\top\widehat{\boldsymbol{W}} = \frac{1}{\phi}\widehat{\boldsymbol{W}}\boldsymbol{H}_{\widehat{W}}(\lambda).\end{aligned}$$

That is, this perturbation scheme is related to the generalized leverage of $\widehat{\boldsymbol{a}}_\lambda$ (Wei et al. 1998), $\boldsymbol{GL}(\widehat{\boldsymbol{a}}_\lambda) = \partial\widehat{\boldsymbol{g}}_\lambda/\partial\boldsymbol{Y}^\top = \boldsymbol{H}_{\widehat{W}}(\lambda)$. In fact, for a fixed or known ϕ, it is possible to study the influence of extreme observations on their own fitted values using the index plot of $B_i(\boldsymbol{a}) \propto \widehat{h}_{ii}(\lambda)$, $i = 1, \ldots, n$ where $\widehat{h}_{ii}(\lambda)$ denotes the ith diagonal element of the $\boldsymbol{H}_{\widehat{W}}(\lambda)$ matrix.

3.2.3 Scale perturbation on the smoothing parameter selection

To evaluate the effect of outlying observations on the smoothing parameter selection, we consider the perturbation scheme defined in (18), In this way, the perturbed WGCV criterion assumes the form

$$V(\lambda, \boldsymbol{\omega}) = \frac{RSS_{\widehat{W}}(\lambda, \boldsymbol{\omega})/n}{\{\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}_{\widehat{W}}(\lambda, \boldsymbol{\omega}))/n\}^2}, \tag{19}$$

where $RSS_{\widehat{W}}(\lambda, \boldsymbol{\omega}) = \|\mathrm{diag}^{-1/2}(\boldsymbol{\omega})\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{I} - \boldsymbol{H}_{\widehat{W}}(\lambda, \boldsymbol{\omega}))\boldsymbol{Y}\|^2$, and

$$\boldsymbol{H}_{\widehat{W}}(\lambda, \boldsymbol{\omega})) = \boldsymbol{B}(\boldsymbol{B}^\top\widehat{\boldsymbol{W}}\mathrm{diag}^{-1}(\boldsymbol{\omega})\boldsymbol{B} + \lambda\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{B}^\top\widehat{\boldsymbol{W}}.$$

Note that the first term of $\partial\widehat{\lambda}(\boldsymbol{\omega})/\partial\boldsymbol{\omega}$ given in Eq. (17) is a scalar that can be ignored, thus

$$\boldsymbol{h}_{\max}(V) \propto \frac{\partial^2 V(\lambda, \boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \lambda}\Big|_{\omega=\omega_0, \lambda=\widehat{\lambda}}.$$

Based on (19), for the smoothing parameter selection with the scale perturbation scheme, the following proposition gives the specific form of the direction of largest local curvature.

Proposition 4 *For the smoothing parameter selection procedure based on the WGCV criterion, under the scale perturbation, the second derivative of $V(\lambda, \boldsymbol{\omega})$ with respect to $\boldsymbol{\omega}$ and λ evaluated at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and $\lambda = \widehat{\lambda}$ can be expressed as*

$$\begin{aligned}
\frac{\partial^2 V(\lambda, \boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \lambda}\Big|_{\omega=\omega_0, \lambda=\widehat{\lambda}} &= -\frac{3}{c}\operatorname{tr}(\boldsymbol{G}\widehat{\boldsymbol{W}})\frac{\partial V(\lambda, \boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\Big|_{\omega=\omega_0, \lambda=\widehat{\lambda}} \\
&\quad + \frac{1}{nc^3}\Big\{\frac{1}{n}\operatorname{tr}(\boldsymbol{G}\widehat{\boldsymbol{W}})\operatorname{diag}(\boldsymbol{e})\widehat{\boldsymbol{W}}(\boldsymbol{I} - 2\widehat{\boldsymbol{H}})\boldsymbol{e} \\
&\quad + c\big[\operatorname{diag}(\boldsymbol{e})\widehat{\boldsymbol{W}}(\boldsymbol{I} - 2\widehat{\boldsymbol{H}})\boldsymbol{G}\widehat{\boldsymbol{W}}\boldsymbol{Y} \\
&\quad + 2\operatorname{diag}(\boldsymbol{e})\widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}}\boldsymbol{e} + \operatorname{diag}(\boldsymbol{G}\widehat{\boldsymbol{W}}\boldsymbol{Y})\widehat{\boldsymbol{W}}(\boldsymbol{I} - 2\widehat{\boldsymbol{H}})\boldsymbol{e}\big] \\
&\quad - \frac{2}{n}\big[RSS_{\widehat{W}}(\widehat{\lambda}, \boldsymbol{\omega}_0)\operatorname{dg}((\boldsymbol{I} - 2\widehat{\boldsymbol{H}})\boldsymbol{G}\widehat{\boldsymbol{W}})\mathbf{1}_n \\
&\quad - 2\operatorname{dg}((\boldsymbol{I} - \widehat{\boldsymbol{H}})\widehat{\boldsymbol{H}})\mathbf{1}_n\boldsymbol{e}^\top\widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}}\boldsymbol{Y}\big]\Big\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial V(\lambda, \boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\Big|_{\omega=\omega_0, \lambda=\widehat{\lambda}} &= \frac{1}{nc^3}\Big\{c\operatorname{diag}(\boldsymbol{e})\widehat{\boldsymbol{W}}(\boldsymbol{I} - 2\widehat{\boldsymbol{H}})\boldsymbol{e} \\
&\quad + \frac{2}{n}RSS_{\widehat{W}}(\widehat{\lambda}, \boldsymbol{\omega}_0)\operatorname{dg}((\boldsymbol{I} - \widehat{\boldsymbol{H}})\widehat{\boldsymbol{H}})\mathbf{1}\Big\} \qquad (20)
\end{aligned}$$

with $\operatorname{dg}(\boldsymbol{Z}) = \operatorname{diag}(z_{11}, \ldots, z_{nn})$ *for* $\boldsymbol{Z} = (z_{ij})$ *a square matrix of order* $n \times n$*, while* $c = 1 - \operatorname{tr}(\widehat{\boldsymbol{H}})/n$*,* $\widehat{\boldsymbol{H}} = \boldsymbol{H}_{\widehat{W}}(\widehat{\lambda})$*,* $\boldsymbol{G} = \boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{K}^\top\boldsymbol{K}\boldsymbol{S}^{-1}\boldsymbol{B}^\top$ *where* $\boldsymbol{S} = \boldsymbol{B}^\top\widehat{\boldsymbol{W}}\boldsymbol{B} + \widehat{\lambda}\boldsymbol{K}^\top\boldsymbol{K}$*,* $\boldsymbol{e} = (\boldsymbol{I} - \widehat{\boldsymbol{H}})\boldsymbol{Y}$ *and* $\mathbf{1}_n = (1, \ldots, 1)^\top$ *denote an n-dimensional vector of ones.*

3.2.4 Response perturbation on the smoothing parameter selection

To perturb the response variable we consider $\boldsymbol{Y}(\boldsymbol{\omega}) = \boldsymbol{Y} + \boldsymbol{\omega}$, with $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ where the vector of null perturbation is $\boldsymbol{\omega}_0 = \mathbf{0}$. The perturbed WGCV criterion assumes the form

$$V(\lambda, \boldsymbol{\omega}) = \frac{RSS_{\widehat{W}}(\lambda, \boldsymbol{\omega})/n}{\{\operatorname{tr}(\boldsymbol{I} - \boldsymbol{H}_{\widehat{W}}(\lambda))/n\}^2},$$

where $RSS_{\widehat{W}}(\lambda, \boldsymbol{\omega}) = \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{I} - \boldsymbol{H}_{\widehat{W}}(\lambda))\boldsymbol{Y}(\boldsymbol{\omega})\|^2$. The following proposition provides an explicit expression for $\partial^2 V(\widehat{\lambda}, \boldsymbol{\omega}_0)/\partial\boldsymbol{\omega}\partial\lambda$.

Proposition 5 *Let $\widehat{\lambda}$ be the selected value of the smoothing parameter according the procedure described in Sect. 2.2. For the response perturbation scheme we have*

$$\boldsymbol{h}_{\max} \propto \frac{\partial^2 V(\lambda, \boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \lambda}\Big|_{\omega=\omega_0, \lambda=\widehat{\lambda}}$$
$$= \frac{2}{nc^2}\left\{(\boldsymbol{I}-\widehat{\boldsymbol{H}})^\top \widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}} + \widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}}(\boldsymbol{I}-\widehat{\boldsymbol{H}}) - \frac{2}{nc}\operatorname{tr}(\boldsymbol{G}\widehat{\boldsymbol{W}})(\boldsymbol{I}-\widehat{\boldsymbol{H}})^\top \widehat{\boldsymbol{W}}(\boldsymbol{I}-\widehat{\boldsymbol{H}})\right\}\boldsymbol{Y},$$

where $c = 1 - \operatorname{tr}(\widehat{\boldsymbol{H}})/n$, $\widehat{\boldsymbol{H}} = \boldsymbol{H}_{\widehat{W}}(\widehat{\lambda})$ *and* $\boldsymbol{G} = \boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{K}^\top \boldsymbol{K}\boldsymbol{S}^{-1}\boldsymbol{B}^\top$ *with* $\boldsymbol{S} = \boldsymbol{B}^\top \widehat{\boldsymbol{W}}\boldsymbol{B} + \widehat{\lambda}\boldsymbol{K}^\top \boldsymbol{K}$.

As we shall see in the confirmatory analysis for life expectancy data, outliers can be extremely influential on the selection of the smoothing parameter (see Table 2). Thus, these perturbation schemes applied on the weighted GCV criterion allow us to note that it is necessary that both, the fitting procedure as the smoothing selection technique

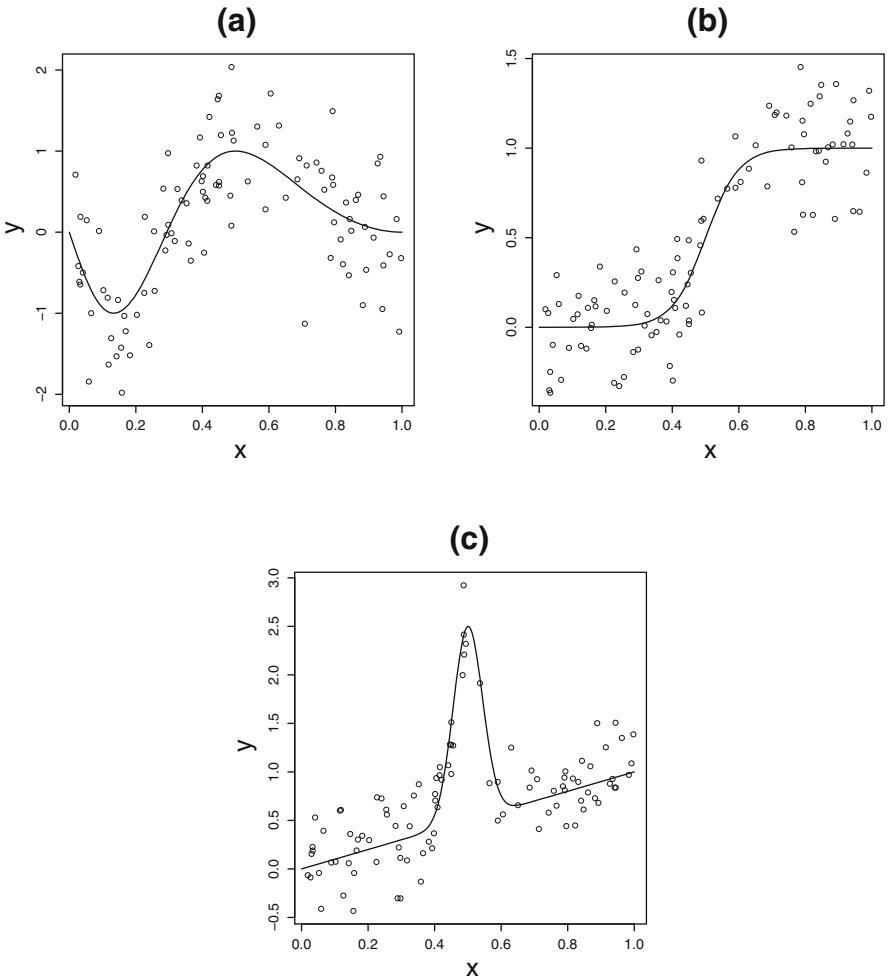

Fig. 1 Typical dataset for: **a** Cantoni and Ronchetti (2001), **b** logistic and **c** "bump" test functions

should be based on a robust approach. Moreover, the influence measure $\boldsymbol{h}_{\max}$ obtained in Propositions 4 and 5 generalizes the work of Thomas (1991).

4 Numerical experiments

In this section, we evaluate the performance of the proposed methodology through a simulation study and the analysis of life expectancy data introduced by Leinhardt and Wasserman (1979). Additional experiments are reported in Appendices C and D from the supplementary material.

4.1 Simulation study

For our simulation study, we considered the model $Y = g(x) + \sigma\epsilon$, with the following test functions:

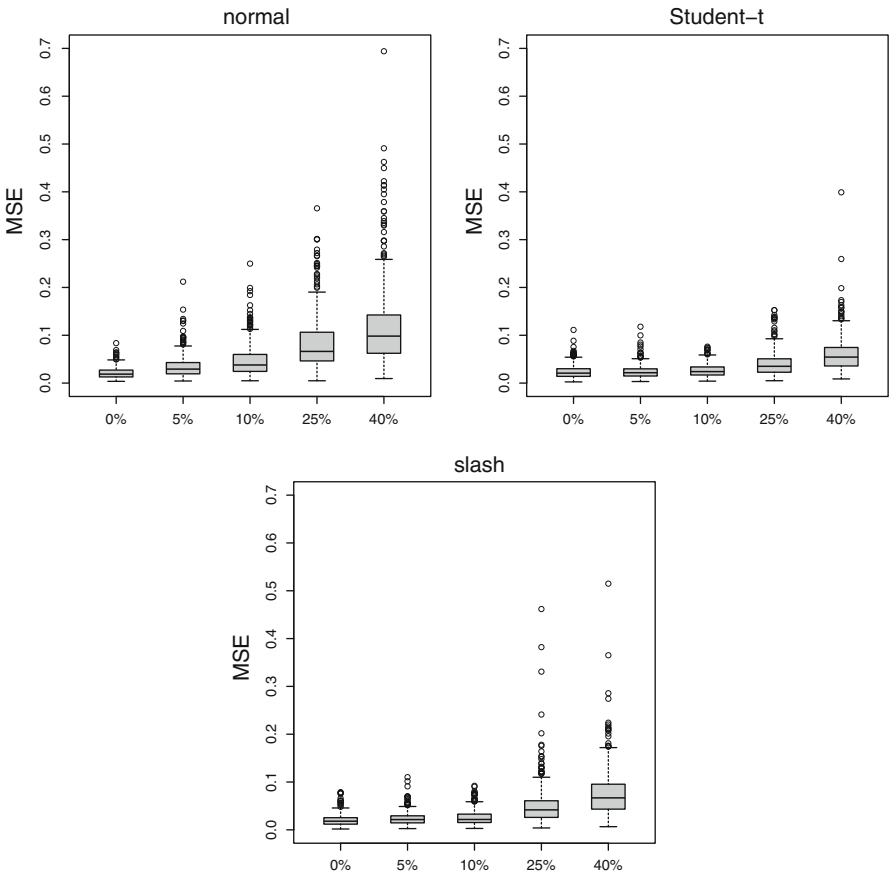

Fig. 2 *Boxplots* of MSE under several contamination levels (δ) and variance inflation factor $\gamma = 4$ considering the three distributional assumptions for the Cantoni and Ronchetti (2001) test function

$$
\begin{aligned}
g_1(x) &= \sin(2\pi(1-x)^2), && \sigma = 0.5,\\
g_2(x) &= \frac{1}{1+\exp(-20(x-1/2))}, && \sigma = 0.2,\\
g_3(x) &= x + 2\exp(-(16(x-1/2))^2), && \sigma = 0.3.
\end{aligned}
$$

Function g_1 was studied by Cantoni and Ronchetti (2001), Lee and Oh (2007) and Tharmaratnam et al. (2010), while functions g_2 (logistic) and g_3 ("bump") where considered by Ruppert (2002) is his simulation study. For each of these functions $M = 500$ datasets were generated. The sample size in each case was $n = 100$ with design points $x_1, \ldots, x_n$ generated independently from the uniform distribution $\mathcal{U}(0, 1)$, the random disturbances $\{\epsilon_i\}$ were generated from the contaminated normal distribution

$$(1-\delta)\mathcal{N}(0,1) + \delta\mathcal{N}(0,\gamma),$$

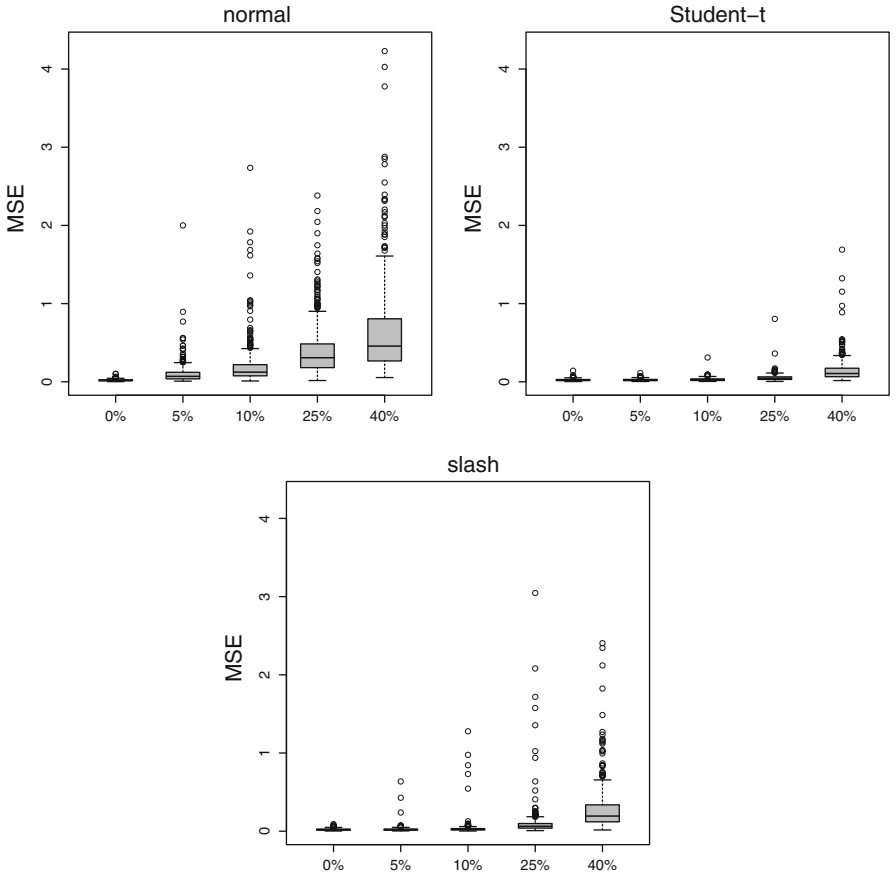

Fig. 3 *Boxplots* of MSE under several contamination levels (δ) and variance inflation factor $\gamma = 10$ considering the three distributional assumptions for the Cantoni and Ronchetti (2001) test function

for $\delta = 0, 5, 10, 25, 40\,\%$ and $\gamma = 2, 4, 10$. We applied P-splines using B-splines of third degree and second order of penalty, with selected knots dividing the domain of x into 20 segments of equal width. The normal, slash and Student t distributions, associated with the $\mathcal{H}$ distribution following point mass at τ_i, Beta and Gamma, respectively, were considered. The degrees of freedom of the slash and Student t were fixed at 2 and 4, respectively. To gain more insights into the performance of the proposed procedure, the mean squared error (MSE) for each simulated dataset j ($j = 1, \ldots, M$), was calculated as

$$MSE_j = \frac{1}{n}\sum_{i=1}^{n}(g(x_i) - \widehat{g}_j(x_i))^2, \qquad j = 1, \ldots, M,$$

where the smoothing parameter λ was chosen according to the strategy outlined in Sect. 2.2. Figure 1 presents a typical dataset for the case in which the data have not been contaminated. As well, the real underlying function is presented for the three test functions considered.

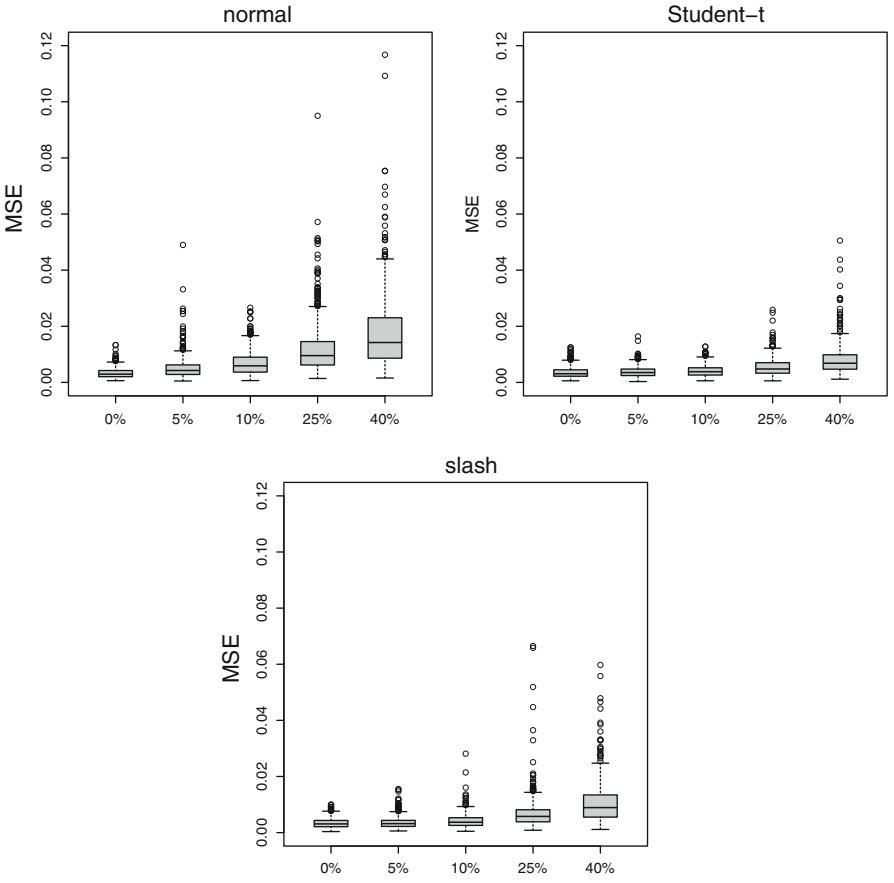

Fig. 4 *Boxplots* of MSE under several contamination levels (δ) and variance inflation factor $\gamma = 4$ considering the three distributional assumptions for the logistic test function

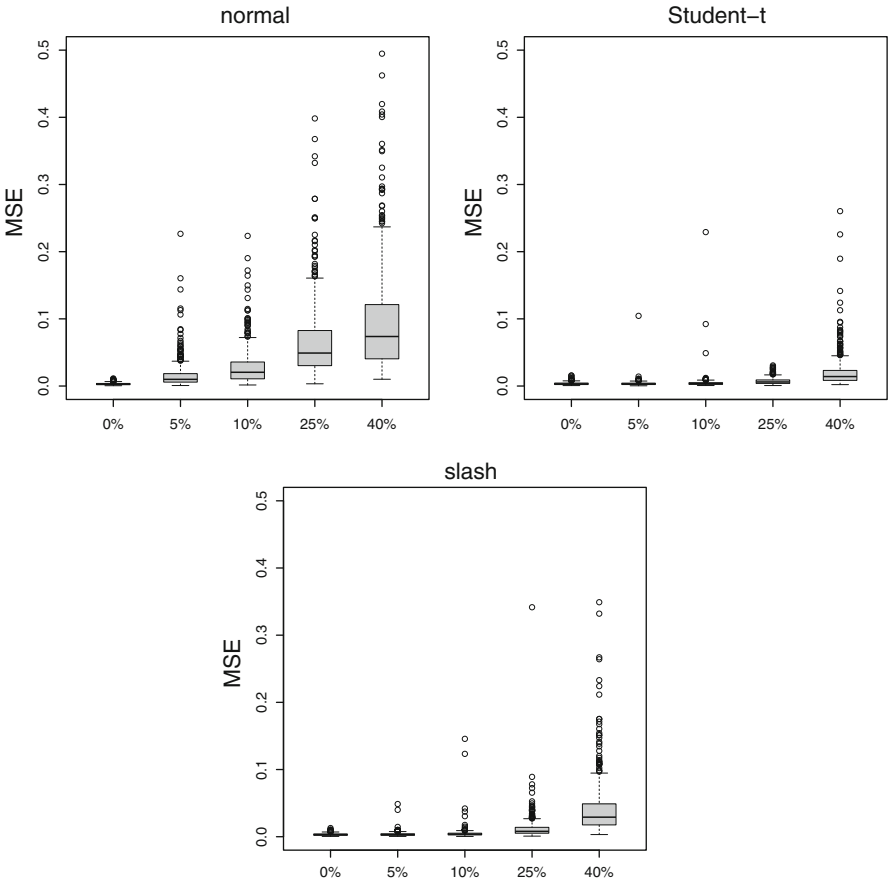

Fig. 5 *Boxplots* of MSE under several contamination levels (δ) and variance inflation factor $\gamma = 10$ considering the three distributional assumptions for the logistic test function

The results of the simulation study are presented in Figs. 2–7. The results for $\gamma = 2$ are omitted because they are similar to those obtained with $\gamma = 4$. As expected, when there is no contamination, the estimated curves using heavy-tailed distributions are essentially equivalent to those obtained under Gaussian errors. However, as the percentage of contamination increases the estimation under normality worsens, while with heavy-tailed distributions the adjustment remains robust against outlying observations. An interesting result in relation to the g_3 "bump" function is presented in Figs. 6 and 7, where it is evident that the protection against outliers offered by the use of heavy-tailed distributions improves only for severe contaminations.

4.2 Life expectancy data

To illustrate the estimation procedure and influence diagnostics described in Sects. 2 and 3, we considered the life expectancy dataset introduced by Leinhardt and Wasserman (1979), who reported on per capita income in US dollars and life expectancy for

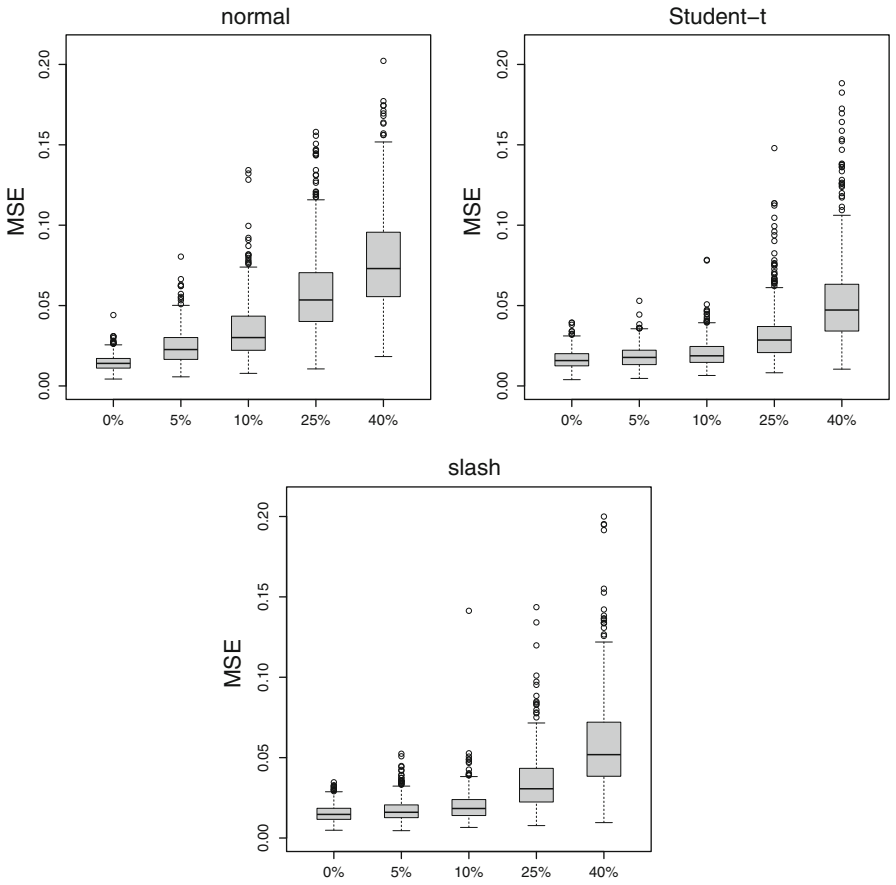

Fig. 6 *Boxplots* of MSE under several contamination levels (δ) and variance inflation factor $\gamma = 4$ considering the three distributional assumptions for the "bump" test function

101 countries in 1979. Thomas (1991) studied the local influence on the GCV criterion in smoothing splines and identified observations 9, 15 and 27 as the most influential using a scale perturbation scheme.

Figure 8 and Table 1 display the results of the fitted model considering normal, slash and Student t distributions. For this dataset, we observe that the estimation procedure under distributions with heavier tails than the normal produce an fitted model that is insensitive to outlying observations. Although there are differences in the estimated values for the degrees of freedom ($\widehat{\nu}$), the smoothing parameter ($\widehat{\lambda}$) and the WGCV criterion, it can be noticed that the estimated curves $\widehat{g}$ for the slash and Student t models are quite similar. It should be stressed that the routine `heavyPS` from the heavy package requires less than a tenth of a second to carry out the computations on an iMac 2,12 Intel Quad Core i5 at 3.1 GHz and 16 GB of RAM.

It is possible to identify outliers in a simple manner by considering the index plot of Mahalanobis distances $D_i^2(\boldsymbol{\theta}) = (Y_i - \boldsymbol{b}_i^\top \boldsymbol{a})^2/\phi$, $i = 1, \ldots, n$ (Lange and

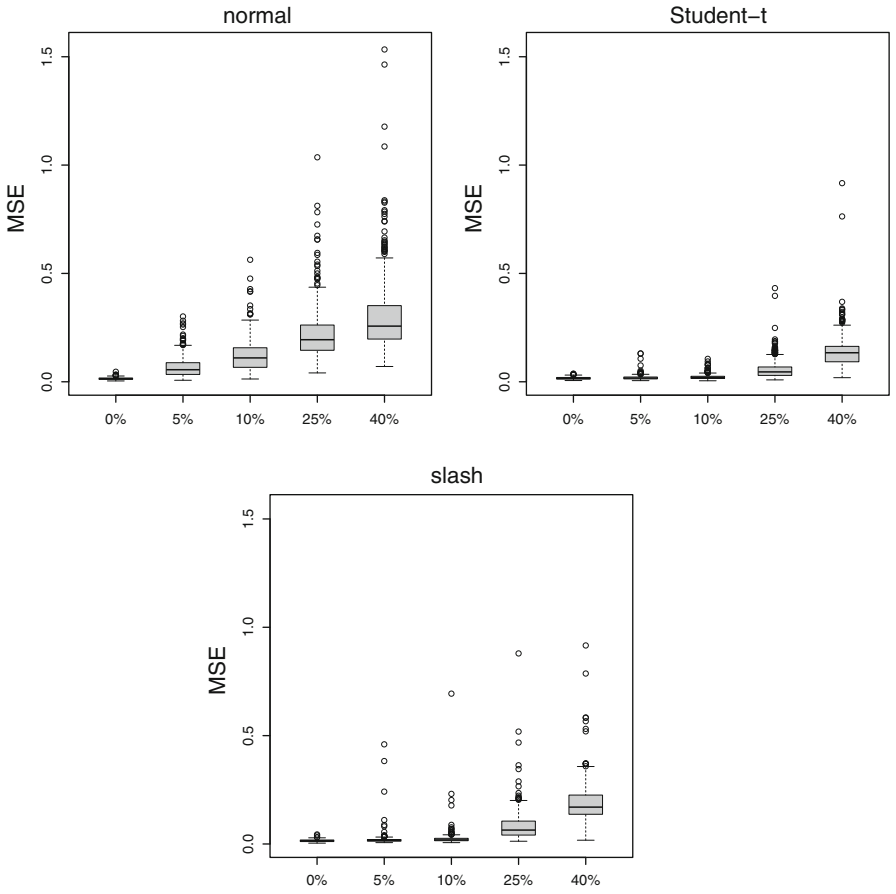

Fig. 7 *Boxplots* of MSE under several contamination levels (δ) and variance inflation factor $\gamma = 10$ considering the three distributional assumptions for the "bump" test function

Sinsheimer 1993). Under normality, $D_i^2(\boldsymbol{\theta})$ follows a Chi-square distribution with one degree of freedom. We use the quantile value $\chi_1^2(\xi)$ with $\xi = 0.975$ to obtain the cutoff shown in the graph in the first panel of Fig. 9. This suggests that under normal errors, observations 49, 58 and 93 are outliers. The other panels in Fig. 9 indicate that when distributions with heavier tails than the normal are used the methodology allows the accommodation of outlying observations (compare with Fig. 8) by attributing small weights in the estimation procedure. Indeed this property is related with the influence function of the PML estimation defined by Eq. (5) (see Butler et al. 1990; Lucas 1997). The weights associated with the normal distribution ($\widehat{\tau}_i = 1$, for $i = 1, \ldots, n$) are indicated in these panels as a dotted line.

To identify influential observations in the life expectancy dataset we applied the influence diagnostic methods proposed in this work. Figure 10 presents the index plot for the generalized Cook distances, GC_i^1, for $i = 1, \ldots, n$, considering the three fitted models. Observations 25 and 27 exercise a strong influence on the PML estimates with

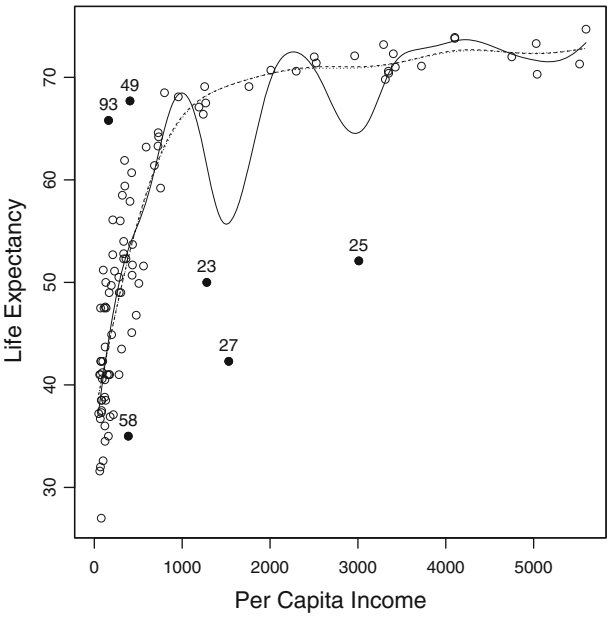

Fig. 8 Life expectancy data with fitted curve under three distributional assumptions: (*solid line*) normal, (*dashed line*) Student t and (*dots*) slash models

Table 1 Estimation summary for the life expectancy data under three fitted models

Model	$\widehat{\nu}$	$\widehat{\lambda}$	$\ell_\lambda(\widehat{\boldsymbol{\theta}}; \boldsymbol{Y}_{\text{obs}})$	WGCV	Iterations	Time (s)
Normal	–	0.1021	−325.4395	48.0393	3	0.002
Slash	1.0966	2.6916	−327.9674	12.8262	104	0.080
Student t	2.5875	4.1905	−326.8620	18.9559	79	0.071

slightly less influence for observations 23 and 35. The influence of those observations decreases considerably when heavy-tailed distributions are used such as slash and Student t distributions. When we consider the assessment of local influence applied to the complete-data-penalized log-likelihood function (Figures 11 and 12) we observe that under normality, observations 23, 25, 27, 58 and 93 are influential against the scale perturbation scheme, while the response perturbation scheme indicates that the fitted model is particularly sensitive when observations 27, 58, and 93 are perturbed. The group of highlighted points in the center panel may be because a property related to the weights definition in the Student t distribution (Kent et al. 1994). It is evident that this influence decreases when we use distributions with heavier tails than the normal.

Figure 13 presents the influence graphs for the scale perturbation scheme applied to the weighted cross-validation procedure. Under errors normally distributed we identify that observations 9, 15 and 27 exercise a strong influence on the selection of the smoothing parameter (see also Thomas 1991). Thus, our results generalize the ones

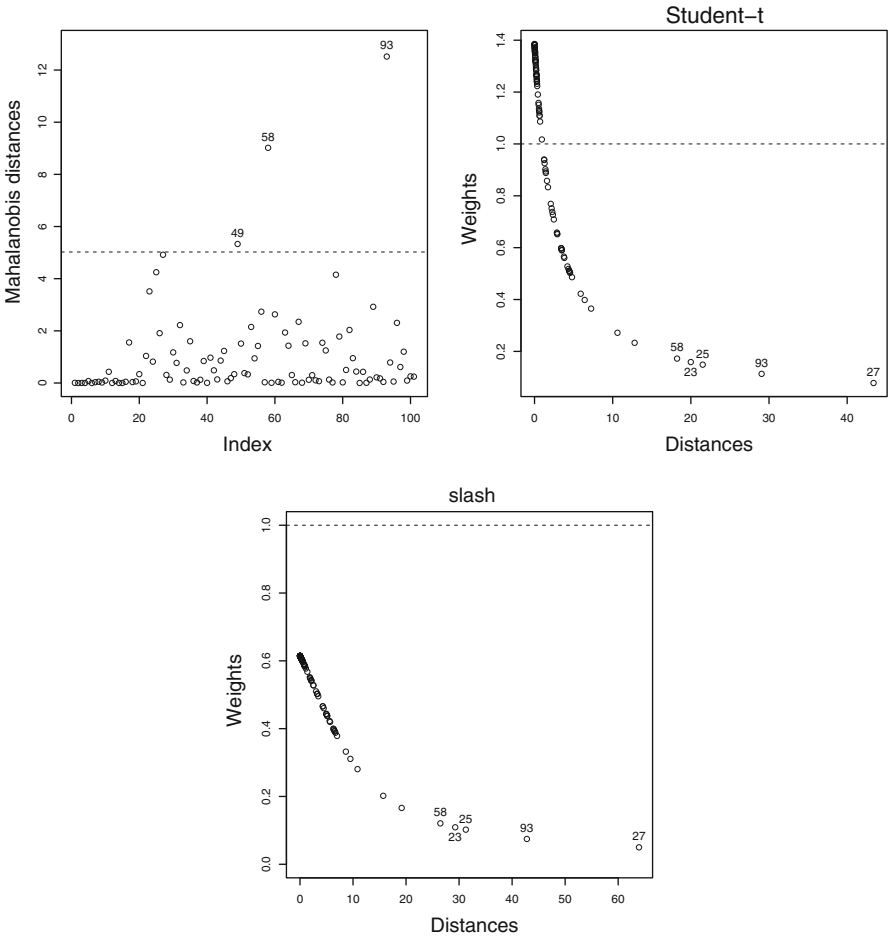

Fig. 9 *Index plot* of the Mahalanobis distances under normality and estimated weights versus distances for the Student t and slash models

reported by Thomas (1991). Again it can be appreciated that the estimation procedure considering heavy-tailed distributions is an effective approach to accommodate outliers.

With the objective of investigating the sensitivity in the selection of the smoothing parameter against outlying observations, a confirmatory study was conducted that consisted of dropping observations from the dataset and obtaining an estimate of λ by minimizing the WGCV criterion. The results were compared with the original estimate. Table 2 presents the estimates and percentages of relative change for each of the fitted models. This analysis reveals the extreme sensitivity in the selection of the smoothing parameter under normal errors. In fact, the highest percentage change (8,257 %) occurs when observations 23, 25 and 27 are removed. The results evidence the ability of SMN distributions to reduce the influence of extreme observations. The stability that can be noted in the estimation of the degrees of

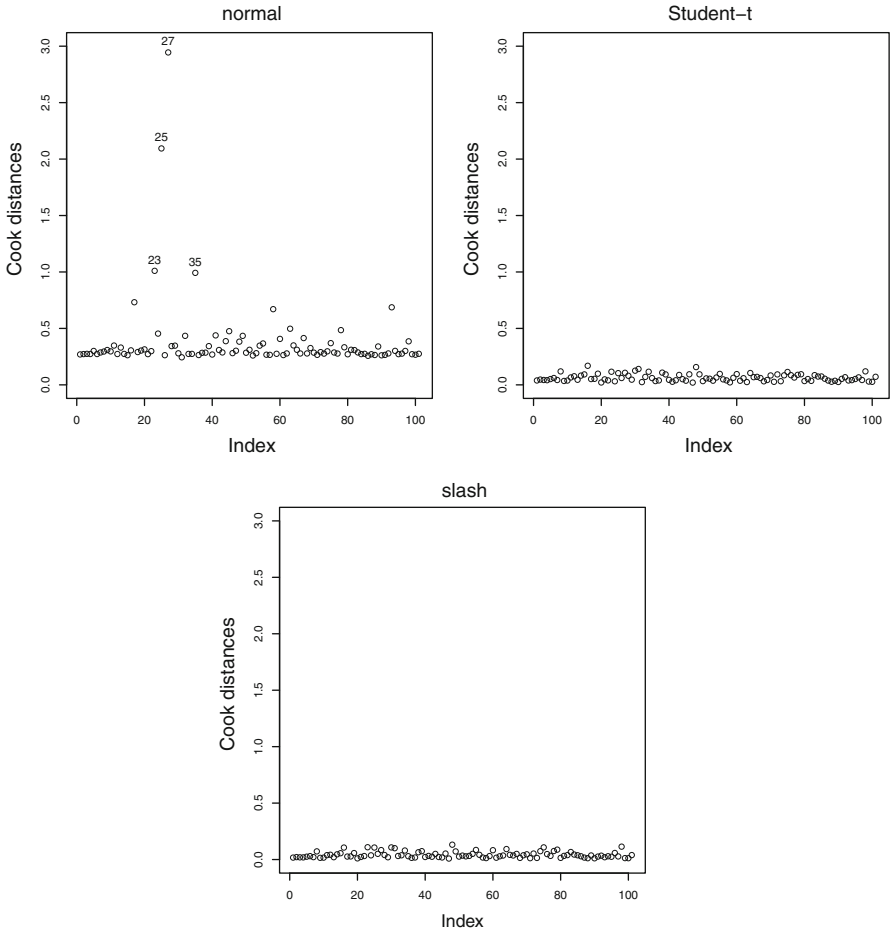

Fig. 10 *Index plots* of generalized Cook distances

freedom for slash and Student t distributions reveals appropriate protection against outliers.

5 Discussion

This work describes the estimation of parameters and the influence diagnostics in smoothing via penalized splines considering the class of scale mixtures of normal distributions. It also addresses the resistant selection of the parameter that controls the smoothness of the fitted curve. The results of the numerical experiments highlight the ability of the proposed procedure to accommodate outlying data. The estimation procedure can be seen as an alternative approach to the works of Cantoni and Ronchetti (2001), Lee and Cox (2009) and Ibacache-Pulgar and Paula (2011). It should be stressed that, using the Laplace distribution (Phillips, 2002) our proce-

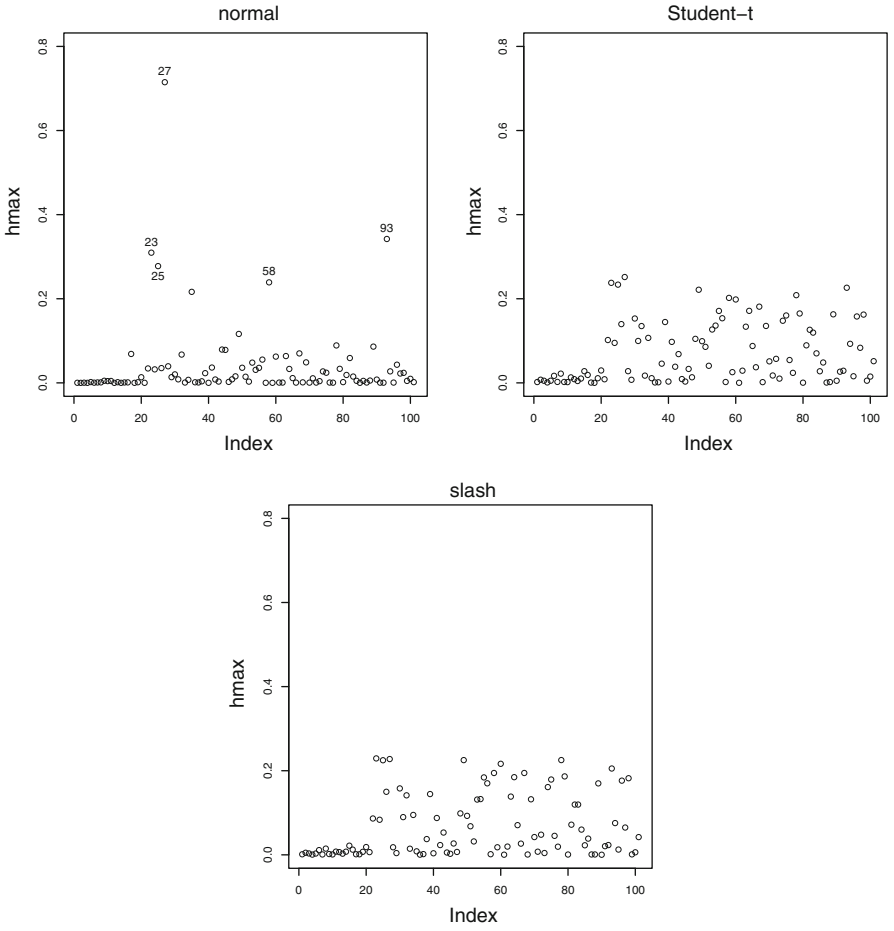

Fig. 11 *Index plots* of $\boldsymbol{h}_{\max}$ under scale perturbation on the penalized log-likelihood function

dure can tackle the median (L_1) nonparametric regression (Koenker et al. 1994). The numerical implementation of the estimation procedure using a penalized EM algorithm is simple and computationally efficient. The routines developed are available in the R heavy package (Osorio 2014). Although we have not seen this situation in our numerical experiments, Gu (1992) presented a discussion in which a series of authors indicated that the procedure delineated in Sect. 2.2 cannot reach convergence. The author is currently working on an alternative form to carry out the selection of the smoothing parameter whose convergence is assured.

The proposed methodology can be seen as an M-estimation procedure (see, for instance Maronna 1976; Lange et al. 1989), thus we can expect that the procedure performs well for variance inflation models (Cook et al. 1982) like the one used in the Monte Carlo simulation study. In an additional simulation study (see supplementary material) we considered an asymmetric contamination scheme varying the percentage

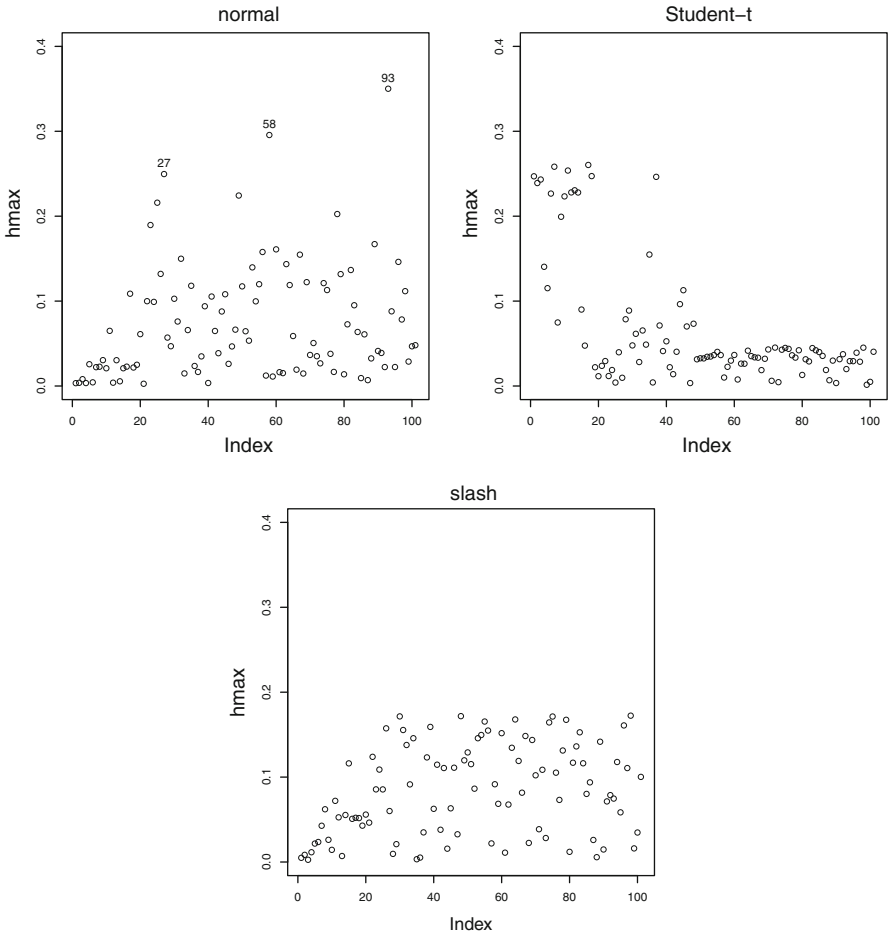

Fig. 12 *Index plots* of $\boldsymbol{h}_{\max}$ under response perturbation on the penalized log-likelihood function

of outliers (0, 5, 10, 25 and 40 %). This kind of extreme contamination reveals that the statistical modeling using distributions with heavier tails than the normal one is not the panacea for all robustness problems (Lange et al. 1989). It is interesting to note that the penalized EM estimation under heavy-tailed distributions produces curve estimates very similar to those reported by Tharmaratnam et al. (2010) when the shape parameters are fixed at very small values. However, models derived under heavy-tailed symmetric distributions, still can be vulnerable to extreme and influential observations. The role and definition of outlying and influential observations in models with heavier tails than the normal one have not been completely studied. In our opinion an avenue for new developments is to use the mean-shift outlier model (Wei and Shih 1994).

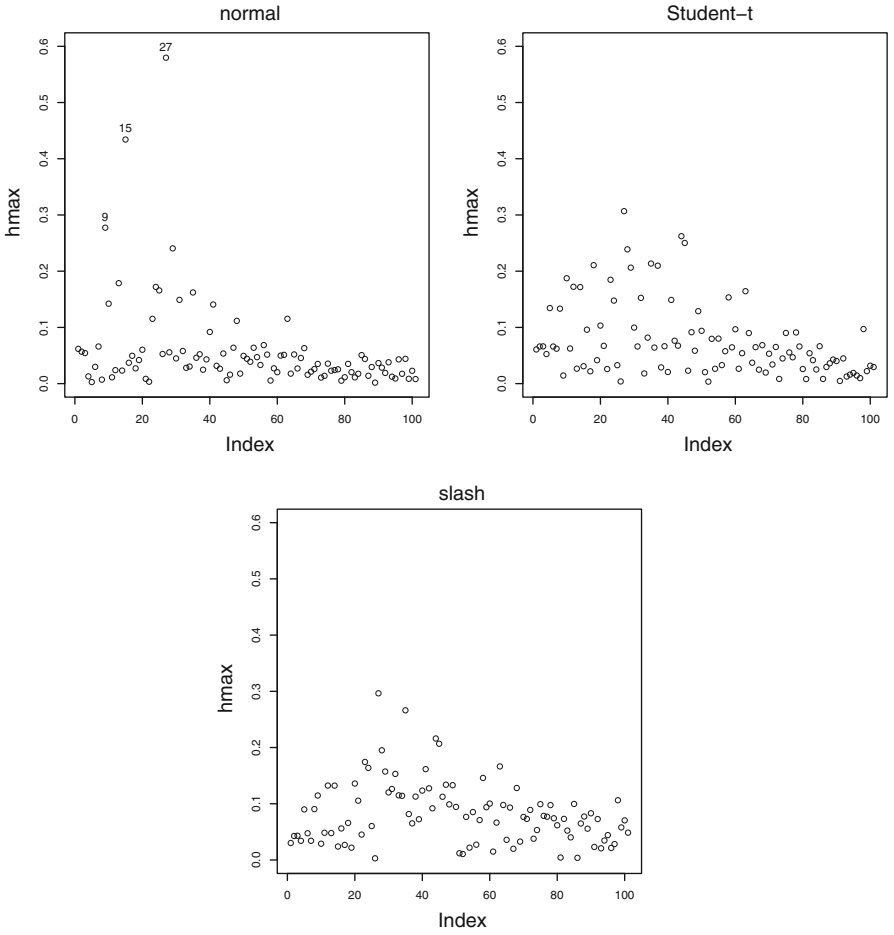

Fig. 13 *Index plots* of $\boldsymbol{h}_{\max}$ under scale perturbation on the smoothing parameter selection

Explicit expressions have been developed for the necessary matrices required to diagnose influence considering case deletion techniques and the local influence method. Interestingly, for the example with real data, all the diagnostic techniques yielded complementary results. The evaluation of influence in this work extends the earlier results of, for example, Eubank and Gunst (1986); Kim (1996) and Wei (2004), who considered the deletion methodology, while the study of local influence generalizes the results of Thomas (1991). We plan to develop an R package to implement the influence diagnostics presented in this work, as a complement to the heavy package.

The results developed in this work can easily be adapted to the context of the ridge regression, in which the proposed methodology extends the works of Walker and Birch (1988); Billor and Loynes (1999) and Shi and Wang (1999). It is planned to extend the parameter estimation as well as the influence assessment considering distributions

Table 2 Selection of smoothing parameter (with percentage change) for the three fitted models

Obs. excluded	Normal		Slash			Student t		
	$\widehat{\lambda}$	[Change (%)]	$\widehat{\lambda}$	[Change (%)]	$\widehat{\nu}$	$\widehat{\lambda}$	[Change (%)]	$\widehat{\nu}$
None	0.1021	–	2.6916	–	1.0966	4.1905	–	2.5875
23	0.1356	33	3.0337	13	1.1663	4.5224	8	2.7612
25	2.7099	2,553	2.6519	−1	1.1673	4.0301	−4	2.7524
27	5.6828	5,464	3.1738	18	1.2792	4.6736	12	3.0603
58	0.0787	−22	2.4344	−10	1.1484	4.0399	−4	3.4976
93	0.0805	−21	2.7566	2	1.2137	4.1696	−1	2.8930
9,15	0.0305	−70	2.5415	−6	1.0647	4.6219	10	3.7019
25,27	5.9363	5,712	3.0141	12	1.2988	4.1281	−1	2.8900
9,15,27	5.5264	5,311	2.9576	10	1.2180	4.5460	8	3.0576
23,25,27	8.5358	8,257	3.9123	45	1.7378	4.5902	10	3.3215
9,15,23,25,27	8.3551	8,080	3.3817	26	1.4475	5.0877	21	4.1768
23,25,27,58,93	4.8414	4,640	3.7382	39	3.0355	4.5478	9	4.5478

with heavier tails than the normal for semiparametric nonlinear mixed effects models according to the approach proposed by Elmi et al. (2011). To reach this objective may require considering an estimation procedure using a stochastic EM algorithm such as that presented by Meza et al. (2012).

Acknowledgments I would like to thank the reviewers for their constructive comments, which helped to substantially improve this manuscript. I am grateful to Victor Leiva for his careful reading and comments on an earlier version of this paper. I also thank Ronny Vallejos and Patricio Videla for their valuable suggestions. The author was partially supported by Grants CONICYT 791100007 and FONDECYT 1140580.

References

Abramowitz, M., and Stegun, I. A. (1970). *Handbook of mathematical functions*. New York: Dover.

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

Billor, N., and Loynes, R. M. (1999). An application of the local influence approach to ridge regression. *Journal of Applied Statistics*, *26*, 177–183.

Butler, R. J., McDonald, J. B., Nelson, R. D., and White, S. B. (1990). Robust and partially adaptive estimation of regression models. *The Review of Economics and Statistics*, *72*, 321–327.

Cantoni, E., and Ronchetti, E. (2001). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, *11*, 141–146.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B*, *48*, 133–169.

Cook, R. D., and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.

Cook, R. D., Holschuh, N., and Weisberg, S. (1982). A note on an alternative outlier model. *Journal of the Royal Statistical Society, Series B*, *44*, 370–376.

Craven, P., and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, *31*, 377–403.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

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

Eilers, P. H. C., and Marx, B. D. (2010). Splines, knots, and penalties. *WIREs Computational Statistics*, *2*, 637–653. doi:10.1002/wics.125.

Elmi, A., Ratcliffe, S. J., Parry, S., and Guo, W. (2011). A B-spline based semiparametric nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, *20*, 492–509.

Escobar, L. A., and Meeker, W. Q. (1992). Assessing influence in regression analysis with censored data. *Biometrics*, *48*, 507–528.

Eubank, R. L. (1984). The hat matrix for smoothing splines. *Statistics & Probability Letters*, *2*, 9–14.

Eubank, R. L. (1985). Diagnostics for smoothing splines. *Journal of the Royal Statistical Society, Series B*, *47*, 332–341.

Eubank, R. L., and Gunst, R. F. (1986). Diagnostics for penalized least-squares estimators. *Statistics & Probability Letters*, *4*, 265–272.

Fernández, C., and Steel, M. F. J. (1999). Multivariate Student-t regression models: Pitfalls and inference. *Biometrika*, *86*, 153–167.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood. *Journal of the Royal Statistical Society, Series B*, *52*, 443–452.

Gu, C. (1992). Cross-validating non-Gaussian data. *Journal of Computational and Graphical Statistics*, *1*, 169–179.

Gu, C., and Xiang, D. (2001). Cross-validating non-gaussian data: Generalized approximate cross-validation revisited. *Journal of Computational and Graphical Statistics*, *10*, 581–591.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Huber, P. J. (1979). Robust smoothing. In R. L. Launer and G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 33–48). New York: Academic Press.

Ibacache-Pulgar, G., and Paula, G. A. (2011). Local influence for Student-t partially linear models. *Computational Statistics & Data Analysis*, *55*, 1462–1478.

Jamshidian, M. (1999). Adaptive robust regression by using a nonlinear regression program. *Journal of Statistical Software*, *4*(6), 1–25.

Kent, J. T., Tyler, D. E., and Vardi, Y. (1994). A curious likelihood identity for the multivariate t-distribution. *Communications in Statistics: Simulation and Computation*, *23*, 441–453.

Kim, C. (1996). Cook's distance in splines smoothing. *Statistics & Probability Letters*, *31*, 139–144.

Kim, C., Park, B. U., and Kim, W. (2002). Influence diagnostics in semiparametric regression models. *Statistics & Probability Letters*, *60*, 49–58.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, *81*, 673–680.

Lange, K., and Sinsheimer, J. S. (1993). Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, *2*, 175–198.

Lange, K., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association*, *84*, 881–896.

Lee, J. S., and Cox, D. D. (2009). Robust smoothing: Smoothing parameter selection and applications to fluorescence spectroscopy. *Computational Statistics & Data Analysis*, *54*, 3131–3143.

Lee, T. C. M., and Oh, H. S. (2007). Robust penalized regression spline fitting with application to additive mixed modeling. *Computational Statistics*, *22*, 159–171.

Leinhardt, S., and Wasserman, S. S. (1979). Teaching regression: An exploratory approach. *The American Statistician*, *33*, 196–203.

Lin, T. I., and Lee, J. C. (2006). A robust approach to t linear mixed models applied to multiple sclerosis data. *Statistics in Medicine*, *25*, 1397–1412.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics*, *37*, 23–38.

Lucas, A. (1997). Robustness of the Student t based M-estimator. *Communications in Statistics: Theory and Methods*, *26*, 1165–1182.

Manchester, L. (1996). Empirical influence for robust smoothing. *Australian Journal of Statistics*, *38*, 275–290.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, *4*, 51–67.

Mateos, G., and Giannakis, G. B. (2012). Robust nonparametric regression via sparsity control with application to load curve data cleansing. *IEEE Transactions on Signal Processing*, *60*, 1571–1584.
McLachlan, G. L., and Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.
Meza, C., Osorio, F., and De la Cruz, R. (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing*, *22*, 121–139.
Moore, R. J. (1982). Algorithm AS 187: Derivatives of the incomplete gamma integral. *Applied Statistics*, *31*, 330–335.
Oh, H. S., Brown, T., and Charbonneau, P. (2004). Period analysis of variable stars by robust smoothing. *Applied Statistics*, *53*, 15–30.
Oh, H. S., Lee, J., and Kim, D. (2008). A recipe for robust estimation using pseudo data. *Journal of the Korean Statistical Society*, *37*, 63–72.
O'Sullivan, F., Yandell, B. S., and Raynor, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association*, *81*, 96–103.
Osorio, F. (2014). heavy: Package for robust estimation using heavy-tailed distributions. R package version 0.2-35. URL: CRAN.R-project.org/package=heavy.
Phillips, R. F. (2002). Least absolute deviations estimation via the EM algorithm. *Statistics and Computing*, *12*, 281–285.
Pinheiro, J., Liu, C., and Wu, Y. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, *10*, 249–276.
Poon, W., and Poon, Y. S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society, Series B*, *61*, 51–61.
R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org.
Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, *11*, 735–757.
Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.
Shi, L., and Wang, X. (1999). Local influence in ridge regression. *Computational Statistics & Data Analysis*, *31*, 341–353.
Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B*, *47*, 1–52.
Staudenmayer, J., Lake, E. E., and Wand, M. P. (2009). Robustness for general design mixed models using the t-distribution. *Statistical Modelling*, *9*, 235–255.
Tharmaratnam, K., Claeskens, G., Croux, C., and Salibián-Barrera, M. (2010). S-estimation for penalized regression splines. *Journal of Computational and Graphical Statistics*, *19*, 609–625.
Thomas, W. (1991). Influence diagnostics for the cross-validated smoothing parameter in spline smoothing. *Journal of the American Statistical Association*, *86*, 693–698.
Utreras, F. I. (1981). On computing robust splines and applications. *SIAM Journal on Scientific and Statistical Computing*, *2*, 153–163.
Walker, E., and Birch, J. B. (1988). Influence measures in ridge regression. *Technometrics*, *30*, 221–227.
Wei, W. H. (2004). Derivatives diagnostics and robustness for smoothing splines. *Computational Statistics & Data Analysis*, *46*, 335–356.
Wei, W. H. (2005). The smoothing parameter, confidence interval and robustness for smoothing splines. *Journal of Nonparametric Statistics*, *17*, 613–642.
Wei, B. C., and Shih, J. Q. (1994). On statistical models for regression diagnostics. *Annals of the Institute of Statistical Mathematics*, *46*, 267–278.
Wei, B. C., Hu, Y. Q., and Fung, W. K. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistics*, *25*, 25–37.
Xiang, D., and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, *6*, 675–692.
Zhu, H., and Lee, S. Y. (2001). Local influence for incomplete-data models. *Journal of the Royal Statistical Society, Series B*, *63*, 111–126.
Zhu, H., Lee, S. Y., Wei, B. C., and Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika*, *88*, 727–737.