INFLUENCE DIAGNOSTICS FOR ROBUST P-SPLINES USING SCALE MIXTURE OF NORMAL DISTRIBUTIONS

FELIPE OSORIO

Supplementary Material

Appendix A presents details about the computational implementation. Proofs of Propositions 1 to 5 are established in Appendix B. Additional results obtained from a real data analysis and a Monte Carlo simulation study are presented in the Appendices C and D, respectively.

Appendix A. Notes about the computational implementation

Below we describe the computational strategy adopted in the heavy R package that uses the estimation procedure proposed in this work. A set of routines has been written in C in order to accelerate some of the calculations. The implementation is fairly simple and most of the cases has been possible to draw upon the routines from the BLAS library (Lawson et al., 1979), Linpack (Dongarra et al., 1979) and Mathlib, included in the R software (R Core Team, 2014).

The M step of the algorithm, described in Equations (9) and (10) from Section 2.1 can be efficiently computed by considering $\boldsymbol{U} = \boldsymbol{W}^{(k)^{1/2}}\boldsymbol{B}$ and decompose $\boldsymbol{U}$ using a singular value decomposition like $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, where $\boldsymbol{U}$ is $n \times p$ matrix such that $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_p$, $\boldsymbol{D} = \text{diag}(\delta_1, \ldots, \delta_p)$ and $\boldsymbol{V}$ is an orthogonal $p \times p$ matrix. It is important to note that the form of decomposition highlights that it is possible to overwrite the matrix $\boldsymbol{U}$ in order to keep storage to a minimum. A second singular value decomposition is then calculated:

$$\boldsymbol{K}\boldsymbol{V}\boldsymbol{D}^{-1} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{S}^\top,$$

where only $\boldsymbol{R} = \text{diag}(r_1, \ldots, r_p)$ and the orthogonal $p \times p$ matrix $\boldsymbol{S}$ are stored. Let $\boldsymbol{Z} = \boldsymbol{U}^\top\boldsymbol{W}^{(k)^{1/2}}\boldsymbol{Y}$ and $\boldsymbol{c} = \boldsymbol{S}^\top\boldsymbol{Z}$. Thus, we define

$$\boldsymbol{\alpha}_\lambda^{(k+1)} = \boldsymbol{S}(\boldsymbol{I} + \lambda\boldsymbol{R}^2)^{-1}\boldsymbol{c}, \quad \lambda > 0, \tag{A.1}$$

and proceed to calculate the fitted values and residuals for the current fit as

$$\boldsymbol{g}_\lambda^{(k+1)} = \boldsymbol{U}\boldsymbol{\alpha}_\lambda^{(k+1)} \quad \text{and} \quad \boldsymbol{e}_\lambda^{(k+1)} = \boldsymbol{Z} - \boldsymbol{g}_\lambda^{(k+1)},$$

respectively. Hence, Equation (10) adopts the form

$$\phi_\lambda^{(k+1)} = \frac{1}{n}\{\|\boldsymbol{e}_\lambda^{(k+1)}\|^2 + \lambda\|\boldsymbol{R}(\boldsymbol{I} + \lambda\boldsymbol{R}^2)^{-1}\boldsymbol{c}\|^2\}. \tag{A.2}$$

The author was supported by grants CONICYT 791100007 and FONDECYT 1140580.

Using the results above, it is possible to rewrite the WGCV criterion defined in (14) from Section 2.2 for the smoothing parameter selection as

$$V(\lambda) = \frac{1}{n}\frac{\|\boldsymbol{e}_{\lambda}^{(k+1)}\|^2}{(1-\mathsf{edf}/n)^2}, \tag{A.3}$$

where edf represents the effective number of parameters, defined as $\mathsf{edf} = \operatorname{tr}\boldsymbol{H}_W(\lambda)$, which can be calculated efficiently as

$$\mathsf{edf} = \operatorname{tr}(\boldsymbol{I} + \lambda\boldsymbol{R}^2)^{-1} = \sum_{j=1}^{p}\frac{1}{1+\lambda\, r_j^2}.$$

We use the Brent's method (Brent, 1973) to carry out the minimization of $V(\lambda)$ given in (A.3). Finally, once convergences of the algorithm was reached, we take

$$\widehat{\boldsymbol{a}}_{\lambda} = \boldsymbol{V}\boldsymbol{D}^{-1}\widehat{\boldsymbol{\alpha}}_{\lambda}, \quad \widehat{\boldsymbol{g}}_{\lambda} = \widehat{\boldsymbol{W}}^{-1/2}\boldsymbol{U}\widehat{\boldsymbol{\alpha}}_{\lambda}.$$

The estimation procedure according to the algorithm described in this section is available in the `heavyPS` function of `heavy` package.

Appendix B. Proofs of main results

In this appendix we derive the differentials $\mathrm{d}^2_{\theta}\, Q_{\lambda}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$, $\mathrm{d}^2_{\omega\theta}\, Q_{\lambda}(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$ and $\mathrm{d}^2_{\omega\theta}\, V(\lambda,\boldsymbol{\omega})$ for scale and response perturbation schemes. The necessary matrices $\ddot{Q}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$, $\boldsymbol{\Delta}(\boldsymbol{\omega})$ and the vector $\partial^2 V(\lambda,\boldsymbol{\omega})/\partial\lambda\partial\boldsymbol{\omega}^{\top}$ are obtained efficiently using the differentiation method and by applying some identification theorems discussed in Magnus and Neudecker (1999).

Proof of Proposition 1. The complete data penalized log-likelihood function given in Equation (6) from Section 2.1 is

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

Differentiating $Q_{\lambda}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{a}$, we obtain

$$\mathrm{d}_a\, Q_{\lambda}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \frac{1}{\phi}\{(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})^{\top}\widehat{\boldsymbol{W}}\boldsymbol{B} - \lambda\boldsymbol{a}^{\top}\boldsymbol{K}^{\top}\boldsymbol{K}\}\,\mathrm{d}\boldsymbol{a},$$

$$\mathrm{d}^2_a\, Q_{\lambda}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = -\frac{1}{\phi}(\mathrm{d}\boldsymbol{a})^{\top}\{\boldsymbol{B}^{\top}\widehat{\boldsymbol{W}}\boldsymbol{B} + \lambda\boldsymbol{K}^{\top}\boldsymbol{K}\}\,\mathrm{d}\boldsymbol{a},$$

and the differential of $\mathrm{d}_a\, Q_{\lambda}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ with respect to ϕ leads to

$$\mathrm{d}^2_{\phi a}\, Q_{\lambda}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = -\frac{1}{\phi^2}\,\mathrm{d}\phi\{(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})^{\top}\widehat{\boldsymbol{W}}\boldsymbol{B} - \lambda\boldsymbol{a}^{\top}\boldsymbol{K}^{\top}\boldsymbol{K}\}\,\mathrm{d}\boldsymbol{a}.$$

In addition, the differentials of $Q_{\lambda}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ with respect to ϕ are given by

$$\mathrm{d}_{\phi}\, Q_{\lambda}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = -\frac{n}{2\phi}\,\mathrm{d}\phi + \frac{1}{2\phi^2}\{(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})^{\top}\widehat{\boldsymbol{W}}(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a}) + \lambda\boldsymbol{a}^{\top}\boldsymbol{K}^{\top}\boldsymbol{K}\boldsymbol{a}\}\,\mathrm{d}\phi,$$

$$\mathrm{d}^2_{\phi}\, Q_{\lambda}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \frac{n}{2\phi^2}\,\mathrm{d}^2\phi - \frac{1}{\phi^3}\{(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})^{\top}\widehat{\boldsymbol{W}}(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a}) + \lambda\boldsymbol{a}^{\top}\boldsymbol{K}^{\top}\boldsymbol{K}\boldsymbol{a}\}\,\mathrm{d}^2\phi.$$

From the first-order conditions

$$\boldsymbol{B}^{\top}\widehat{\boldsymbol{W}}\boldsymbol{Y} - (\boldsymbol{B}^{\top}\widehat{\boldsymbol{W}}\boldsymbol{B} + \lambda\boldsymbol{K}^{\top}\boldsymbol{K})\widehat{\boldsymbol{a}} = \boldsymbol{0}, \tag{B.1}$$

$$n\widehat{\phi} - (RSS_{\widehat{W}}(\widehat{\boldsymbol{a}}) + \lambda\widehat{\boldsymbol{a}}^{\top}\boldsymbol{K}^{\top}\boldsymbol{K}\widehat{\boldsymbol{a}}) = 0, \tag{B.2}$$

follow immediately that

$$\mathrm{d}_a^2\, Q_\lambda(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})\big|_{\theta=\widehat{\theta}} = -\frac{1}{\widehat{\phi}}(\mathrm{d}\boldsymbol{a})^\top\{\boldsymbol{B}^\top\widehat{\boldsymbol{W}}\boldsymbol{B}+\lambda\boldsymbol{K}^\top\boldsymbol{K}\}\,\mathrm{d}\boldsymbol{a},$$

$$\mathrm{d}_{\phi a}^2\, Q_\lambda(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})\big|_{\theta=\widehat{\theta}} = \boldsymbol{0}, \qquad \mathrm{d}_\phi^2\, Q_\lambda(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})\big|_{\theta=\widehat{\theta}} = -\frac{n}{2\widehat{\phi}^2}\,\mathrm{d}^2\phi.$$

Applying the second identification theorem given in Magnus and Neudecker (1999) we obtain $\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$, and the proof is complete. □

Proof of Proposition 2. The Q-function for the perturbated model introduced in Equation (18) from Subsection 3.2.1 assumes the form

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

where $\boldsymbol{\omega}$ is an n-dimensional perturbation vector, the perturbated model reduces to the postulated model when $\boldsymbol{\omega}_0 = \boldsymbol{1}$.

Taking differentials of the $Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta} = (\boldsymbol{a}^\top,\phi)^\top$, we get

$$\mathrm{d}_a\, Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \frac{1}{\phi}\{(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})^\top\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega})\boldsymbol{B} - \lambda\boldsymbol{a}^\top\boldsymbol{K}^\top\boldsymbol{K}\}\,\mathrm{d}\boldsymbol{a},$$

$$\begin{aligned}\mathrm{d}_\phi\, Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = &-\frac{n}{2\phi}\,\mathrm{d}\phi + \frac{1}{2\phi^2}\{(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})^\top\widehat{\boldsymbol{W}}^{1/2}\,\mathrm{diag}(\boldsymbol{\omega})\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a})\\ &+\lambda\boldsymbol{a}^\top\boldsymbol{K}^\top\boldsymbol{K}\boldsymbol{a}\}\,\mathrm{d}\phi,\end{aligned}$$

using that $\boldsymbol{z}^\top\,\mathrm{diag}(\boldsymbol{\omega}) = \boldsymbol{\omega}^\top\,\mathrm{diag}(\boldsymbol{z})$ and $\mathrm{diag}(\boldsymbol{\omega})\,\mathrm{diag}(\boldsymbol{z}) = \mathrm{diag}(\boldsymbol{z})\,\mathrm{diag}(\boldsymbol{\omega})$ for $\boldsymbol{z}$ an n-dimensional vector, the differentials $\mathrm{d}_a\, Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$ and $\mathrm{d}_\phi\, Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$ can be written as

$$\mathrm{d}_a\, Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \frac{1}{\phi}\{\boldsymbol{\omega}^\top\,\mathrm{diag}(\boldsymbol{\epsilon})\widehat{\boldsymbol{W}}\boldsymbol{B} - \lambda\boldsymbol{a}^\top\boldsymbol{K}^\top\boldsymbol{K}\}\,\mathrm{d}\boldsymbol{a},$$

$$\mathrm{d}_\phi\, Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = -\frac{n}{2\phi}\,\mathrm{d}\phi + \frac{1}{2\phi^2}\{\boldsymbol{\omega}^\top\,\mathrm{diag}(\boldsymbol{\epsilon})\widehat{\boldsymbol{W}}\boldsymbol{\epsilon} + \lambda\boldsymbol{a}^\top\boldsymbol{K}^\top\boldsymbol{K}\boldsymbol{a}\}\,\mathrm{d}\phi,$$

where $\boldsymbol{\epsilon} = \boldsymbol{Y}-\boldsymbol{B}\boldsymbol{a}$. To find $\boldsymbol{\Delta}(\boldsymbol{\omega})$, we differentiate $\mathrm{d}_a\, Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$ and $\mathrm{d}_\phi\, Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\omega}$, thus

$$\mathrm{d}_{\omega a}^2\, Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \frac{1}{\phi}(\mathrm{d}\boldsymbol{\omega})^\top\,\mathrm{diag}(\boldsymbol{\epsilon})\widehat{\boldsymbol{W}}\boldsymbol{B}\,\mathrm{d}\boldsymbol{a}, \tag{B.3}$$

$$\mathrm{d}_{\omega\phi}^2\, Q_\lambda(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \frac{1}{2\phi^2}(\mathrm{d}\boldsymbol{\omega})^\top\,\mathrm{diag}(\boldsymbol{\epsilon})\widehat{\boldsymbol{W}}\boldsymbol{\epsilon}\,\mathrm{d}\phi. \tag{B.4}$$

Applying the second identification theorem given in Magnus and Neudecker (1999) and evaluating (B.3) and (B.4) at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ we obtain $\boldsymbol{\Delta}(\boldsymbol{\omega}_0)$, and the proof is complete. □

Proof of Proposition 3. Under the response perturbation $\boldsymbol{Y}(\boldsymbol{\omega}) = \boldsymbol{Y}+\boldsymbol{\omega}$, the Q-function for the perturbated model is given by

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

where $\boldsymbol{\omega}$ is an n-dimensional perturbation vector and the vector of null perturbation vector is given by $\boldsymbol{\omega}_0 = \boldsymbol{0}$.

The first differential of $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta} = (\boldsymbol{a}^\top, \phi)^\top$ is

$$\mathrm{d}_a\, Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \frac{1}{\phi}\{(\boldsymbol{Y}(\boldsymbol{\omega}) - \boldsymbol{B}\boldsymbol{a})^\top \widehat{\boldsymbol{W}}\boldsymbol{B} - \lambda \boldsymbol{a}^\top \boldsymbol{K}^\top \boldsymbol{K}\}\,\mathrm{d}\boldsymbol{a},$$

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

taking the differential of $\mathrm{d}_a\, Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$ and $\mathrm{d}_\phi\, Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\omega}$ we have

$$\mathrm{d}^2_{\omega a}\, Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \frac{1}{\phi}(\mathrm{d}\boldsymbol{\omega})^\top \widehat{\boldsymbol{W}}\boldsymbol{B}\,\mathrm{d}\boldsymbol{a}, \tag{B.5}$$

$$\mathrm{d}^2_{\omega\phi}\, Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \frac{1}{2\phi^2}(\mathrm{d}\boldsymbol{\omega})^\top \widehat{\boldsymbol{W}}(\boldsymbol{Y}(\boldsymbol{\omega}) - \boldsymbol{B}\boldsymbol{a})\,\mathrm{d}\phi. \tag{B.6}$$

The $\boldsymbol{\Delta}(\boldsymbol{\omega}_0)$ matrix can be obtained by applying the second identification theorem (Magnus and Neudecker, 1999) and evaluating (B.5) and (B.6) at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Thus, the proposition is verified. □

Proof of Proposition 4. The weighted cross-validation criterion under the scale perturbation defined by Equation (19) from Subsection 3.2.3 is given by

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

where

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

and

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

Next, we shall write $RSS_{\widehat{W}}(\lambda, \boldsymbol{\omega}) = RSS(\lambda, \boldsymbol{\omega})$ and $\boldsymbol{H}_{\widehat{W}}(\lambda, \boldsymbol{\omega}) = \boldsymbol{H}(\lambda, \boldsymbol{\omega})$. Let $\mathsf{edf}(\lambda, \boldsymbol{\omega}) = \mathrm{tr}\,\boldsymbol{H}(\lambda, \boldsymbol{\omega})$ be the effective number of parameters for the perturbed model. Taking the differential of $V(\lambda, \boldsymbol{\omega})$ with respect to $\boldsymbol{\omega}$ leads to

$$\begin{aligned}\mathrm{d}_\omega\, V(\lambda, \boldsymbol{\omega}) = \frac{1/n}{(1 - \mathsf{edf}(\lambda, \boldsymbol{\omega})/n)^4}\Big\{&(1 - \mathsf{edf}(\lambda, \boldsymbol{\omega})/n)^2\,\mathrm{d}_\omega\, RSS(\lambda, \boldsymbol{\omega}) \\ &- RSS(\lambda, \boldsymbol{\omega})\,\mathrm{d}_\omega(1 - \mathsf{edf}(\lambda, \boldsymbol{\omega})/n)^2\Big\},\end{aligned} \tag{B.7}$$

is easy to see that,

$$\begin{aligned}\mathrm{d}_\omega\, \boldsymbol{H}(\lambda, \boldsymbol{\omega}) &= -\boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{B}^\top \widehat{\boldsymbol{W}}\,\mathrm{diag}(\mathrm{d}\boldsymbol{\omega})\boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{B}^\top \widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega}) + \boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{B}^\top \widehat{\boldsymbol{W}}\,\mathrm{diag}(\mathrm{d}\boldsymbol{\omega}) \\ &= \boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{B}^\top \widehat{\boldsymbol{W}}\,\mathrm{diag}(\mathrm{d}\boldsymbol{\omega})(\boldsymbol{I} - \boldsymbol{H}(\lambda, \boldsymbol{\omega}))\end{aligned} \tag{B.8}$$

where $\boldsymbol{S} = \boldsymbol{B}^\top \widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega})\boldsymbol{B} + \lambda \boldsymbol{K}^\top \boldsymbol{K}$. Thus, using the properties of the trace operator, we obtain

$$\mathrm{d}_\omega(1 - \mathsf{edf}(\lambda, \boldsymbol{\omega})/n)^2 = -\frac{2}{n}(1 - \mathsf{edf}(\lambda, \boldsymbol{\omega})/n) \times \boldsymbol{1}^\top \mathrm{dg}((\boldsymbol{I} - \boldsymbol{H}(\lambda, \boldsymbol{\omega}))\boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{B}^\top \widehat{\boldsymbol{W}})\,\mathrm{d}\boldsymbol{\omega}. \tag{B.9}$$

Let $\boldsymbol{e}_\omega = (\boldsymbol{I} - \boldsymbol{H}(\lambda, \boldsymbol{\omega}))\boldsymbol{Y}$, direct calculations show that

$$\begin{aligned}\mathrm{d}_\omega RSS(\lambda, \boldsymbol{\omega}) = &-\boldsymbol{Y}^\top(\mathrm{d}_\omega \boldsymbol{H}(\lambda, \boldsymbol{\omega}))^\top \widehat{\boldsymbol{W}}^{1/2}\,\mathrm{diag}(\boldsymbol{\omega})\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{e}_\omega \\ &- \boldsymbol{e}_\omega^\top \widehat{\boldsymbol{W}}^{1/2}\,\mathrm{diag}(\boldsymbol{\omega})\widehat{\boldsymbol{W}}^{1/2}(\mathrm{d}_\omega \boldsymbol{H}(\lambda, \boldsymbol{\omega}))\boldsymbol{Y} \\ &+ \boldsymbol{e}_\omega^\top \widehat{\boldsymbol{W}}^{1/2}\,\mathrm{diag}(\mathrm{d}\boldsymbol{\omega})\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{e}_\omega.\end{aligned} \tag{B.10}$$

Substituting equation (B.8) into (B.10), and using some simple algebra we obtain

$$\mathrm{d}_\omega RSS(\lambda,\boldsymbol{\omega}) = \boldsymbol{e}_\omega^\top(\boldsymbol{I}-2\boldsymbol{H}(\lambda,\boldsymbol{\omega}))^\top\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{e}_\omega)\,\mathrm{d}\boldsymbol{\omega}. \tag{B.11}$$

Therefore, using Equations (B.9) and (B.10) the first differential of $V(\lambda,\boldsymbol{\omega})$ with respect to $\boldsymbol{\omega}$ given in Equation (B.7) can now be written as

$$\begin{aligned}\mathrm{d}_\omega V(\lambda,\boldsymbol{\omega}) = \frac{1/n}{(1-\mathsf{edf}(\lambda,\boldsymbol{\omega})/n)^3}\Big\{&(1-\mathsf{edf}(\lambda,\boldsymbol{\omega})/n)\boldsymbol{e}_\omega^\top(\boldsymbol{I}-2\boldsymbol{H}(\lambda,\boldsymbol{\omega}))^\top\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{e}_\omega)\\ &+\frac{2}{n}RSS(\lambda,\boldsymbol{\omega})\mathbf{1}^\top\mathrm{dg}((\boldsymbol{I}-\boldsymbol{H}(\lambda,\boldsymbol{\omega}))\boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{B}^\top\widehat{\boldsymbol{W}})\Big\}\,\mathrm{d}\boldsymbol{\omega}.\end{aligned} \tag{B.12}$$

Applying the first identification theorem by Magnus and Neudecker (1999) and evaluating (B.12) at $\lambda=\widehat{\lambda}$ and $\boldsymbol{\omega}=\boldsymbol{\omega}_0$ leads to $\partial V(\widehat{\lambda},\boldsymbol{\omega}_0)/\partial\boldsymbol{\omega}$ given in Equation (20) from Subsection 3.2.3

Using matrix differentiation, we have

$$\mathrm{d}_\lambda\,\boldsymbol{H}(\lambda,\boldsymbol{\omega}) = -\boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega})\,\mathrm{d}\lambda, \tag{B.13}$$

where $\boldsymbol{G}=\boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{K}^\top\boldsymbol{K}\boldsymbol{S}^{-1}\boldsymbol{B}^\top$. Thus,

$$\mathrm{d}_\lambda(1-\mathsf{edf}(\lambda,\boldsymbol{\omega})/n)^{-3} = -\frac{3}{n}(1-\mathsf{edf}(\lambda,\boldsymbol{\omega})/n)^{-4}\,\mathrm{tr}(\boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega}))\,\mathrm{d}\lambda, \tag{B.14}$$

and

$$\mathrm{d}_\lambda\,\boldsymbol{e}_\omega = -(\mathrm{d}_\lambda\,\boldsymbol{H}(\lambda,\boldsymbol{\omega}))\boldsymbol{Y} = \boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega})\boldsymbol{Y}\,\mathrm{d}\lambda. \tag{B.15}$$

Therefore, taking the differential of $RSS(\lambda,\boldsymbol{\omega})$ with respect to λ we obtain

$$\mathrm{d}_\lambda\,RSS(\lambda,\boldsymbol{\omega}) = 2\,\boldsymbol{Y}^\top\,\mathrm{diag}(\boldsymbol{\omega})\widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega})\boldsymbol{e}_\omega\,\mathrm{d}\lambda \tag{B.16}$$

Using Equations (B.12)-(B.16), and after some algebra yields to the differential of $\mathrm{d}_\omega V(\lambda,\boldsymbol{\omega})$ with respect to λ which is given by

$$\begin{aligned}\mathrm{d}^2_{\lambda\omega}V(\lambda,\boldsymbol{\omega}) = &-\frac{3}{1-\mathsf{edf}(\lambda,\boldsymbol{\omega})/n}\,\mathrm{tr}(\boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega}))\,\mathrm{d}\lambda\,\mathrm{d}_\omega V(\lambda,\boldsymbol{\omega})\\ &+\frac{1/n}{(1-\mathsf{edf}(\lambda,\boldsymbol{\omega})/n)^3}\Big\{\frac{1}{n}\,\mathrm{tr}(\boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega}))\boldsymbol{e}_\omega^\top(\boldsymbol{I}-2\boldsymbol{H}(\lambda,\boldsymbol{\omega}))^\top\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{e}_\omega)\\ &+(1-\mathsf{edf}(\lambda,\boldsymbol{\omega})/n)\big[\boldsymbol{Y}^\top\,\mathrm{diag}(\boldsymbol{\omega})\widehat{\boldsymbol{W}}\boldsymbol{G}(\boldsymbol{I}-2\boldsymbol{H}(\lambda,\boldsymbol{\omega}))^\top\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{e}_\omega)\\ &+2\boldsymbol{e}_\omega^\top\,\mathrm{diag}(\boldsymbol{\omega})\widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{e}_\omega)+\boldsymbol{e}_\omega^\top(\boldsymbol{I}-2\boldsymbol{H}(\lambda,\boldsymbol{\omega}))^\top\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega})\boldsymbol{Y})\big]\\ &+\frac{4}{n}\boldsymbol{Y}^\top\,\mathrm{diag}(\boldsymbol{\omega})\widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega})\boldsymbol{e}_\omega\mathbf{1}^\top\,\mathrm{dg}((\boldsymbol{I}-\boldsymbol{H}(\lambda,\boldsymbol{\omega}))\boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{B}^\top\widehat{\boldsymbol{W}})\\ &+\frac{2}{n}RSS(\lambda,\boldsymbol{\omega})\mathbf{1}_n^\top\,\mathrm{dg}(\boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{diag}(\boldsymbol{\omega})\boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{B}^\top\widehat{\boldsymbol{W}}\\ &-(\boldsymbol{I}-\boldsymbol{H}(\lambda,\boldsymbol{\omega}))\boldsymbol{G}\widehat{\boldsymbol{W}})\Big\}\,\mathrm{d}\lambda\,\mathrm{d}\boldsymbol{\omega},\end{aligned} \tag{B.17}$$

where the differential $\mathrm{d}_\omega V(\lambda,\boldsymbol{\omega})$ is given in Equation (B.12) and $\mathrm{dg}(\boldsymbol{Z})=\mathrm{diag}(z_{11},\ldots,z_{nn})=\boldsymbol{I}\odot\boldsymbol{Z}$ for $\boldsymbol{Z}=(z_{ij})$ a square matrix of order $n\times n$ with $\odot$ being the hadamard product. Thus, evaluating (B.17) at $\lambda=\widehat{\lambda}$ and $\boldsymbol{\omega}=\boldsymbol{\omega}_0$ and applying the second identification theorem by Magnus and Neudecker (1999) we obtain $\partial^2V(\lambda,\boldsymbol{\omega}_0)/\partial\lambda\partial\boldsymbol{\omega}^\top|_{\lambda=\widehat{\lambda}}$ and the proof is complete. □

Proof of Proposition 5. Under the response perturbation scheme, $\boldsymbol{Y}(\boldsymbol{\omega}) = \boldsymbol{Y} + \boldsymbol{\omega}$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ with $\boldsymbol{\omega}_0 = \mathbf{0}$, the perturbed WGCV criterion assumes the form

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

where

$$RSS_{\widehat{W}}(\lambda, \boldsymbol{\omega}) = \boldsymbol{Y}^\top(\boldsymbol{\omega})(\boldsymbol{I} - \boldsymbol{H}_{\widehat{W}}(\lambda))^\top \widehat{\boldsymbol{W}}(\boldsymbol{I} - \boldsymbol{H}_{\widehat{W}}(\lambda))\boldsymbol{Y}(\boldsymbol{\omega}),$$

and

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

In order to simplify the notation we will write $RSS_{\widehat{W}}(\lambda, \boldsymbol{\omega}) = RSS(\lambda, \boldsymbol{\omega})$ and $\boldsymbol{H}_{\widehat{W}}(\lambda) = \boldsymbol{H}(\lambda)$. The first differential of $V(\lambda, \boldsymbol{\omega})$ with respect to $\boldsymbol{\omega}$ assumes the form

$$\mathrm{d}_\omega V(\lambda, \boldsymbol{\omega}) = \frac{1}{n}\frac{\mathrm{d}_\omega RSS(\lambda, \boldsymbol{\omega})}{\{1 - \mathrm{tr}(\boldsymbol{H}(\lambda))/n\}^2},$$

where

$$\mathrm{d}_\omega RSS(\lambda, \boldsymbol{\omega}) = 2\,\boldsymbol{Y}^\top(\boldsymbol{\omega})(\boldsymbol{I} - \boldsymbol{H}(\lambda))^\top \widehat{\boldsymbol{W}}(\boldsymbol{I} - \boldsymbol{H}(\lambda))\,\mathrm{d}\boldsymbol{\omega}.$$

Let $\mathsf{edf}(\lambda) = \mathrm{tr}\,\boldsymbol{H}(\lambda)$ be the effective number of parameters, thus taking the differential of $\mathrm{d}_\omega V(\lambda, \boldsymbol{\omega})$ with respect to λ yields

$$\begin{aligned}\mathrm{d}^2_{\lambda\omega} V(\lambda, \boldsymbol{\omega}) = {} & \frac{2}{n}\,\mathrm{d}_\lambda(1 - \mathsf{edf}(\lambda, \boldsymbol{\omega})/n)^{-2}\,\boldsymbol{Y}^\top(\boldsymbol{\omega})(\boldsymbol{I} - \boldsymbol{H}(\lambda))^\top \widehat{\boldsymbol{W}}(\boldsymbol{I} - \boldsymbol{H}(\lambda))\,\mathrm{d}\boldsymbol{\omega} \\ & + \frac{2/n}{(1 - \mathsf{edf}(\lambda, \boldsymbol{\omega})/n)^2}\boldsymbol{Y}^\top(\boldsymbol{\omega})\,\mathrm{d}_\lambda[(\boldsymbol{I} - \boldsymbol{H}(\lambda))^\top \widehat{\boldsymbol{W}}(\boldsymbol{I} - \boldsymbol{H}(\lambda))]\,\mathrm{d}\boldsymbol{\omega},\end{aligned} \tag{B.18}$$

Let $\boldsymbol{S} = \boldsymbol{B}^\top \widehat{\boldsymbol{W}}\boldsymbol{B} + \lambda \boldsymbol{K}^\top \boldsymbol{K}$, and using that

$$\mathrm{d}_\lambda \boldsymbol{H}(\lambda) = -\boldsymbol{B}\boldsymbol{S}^{-1}(\mathrm{d}_\lambda \boldsymbol{S})\boldsymbol{S}^{-1}\boldsymbol{B}^\top \widehat{\boldsymbol{W}} = -\boldsymbol{G}\widehat{\boldsymbol{W}}\,\mathrm{d}\lambda,$$

where $\boldsymbol{G} = \boldsymbol{B}\boldsymbol{S}^{-1}\boldsymbol{K}^\top \boldsymbol{K}\boldsymbol{S}^{-1}\boldsymbol{B}^\top$, leads to

$$\begin{aligned}\mathrm{d}_\lambda(\boldsymbol{I} - \boldsymbol{H}(\lambda))^\top \widehat{\boldsymbol{W}}(\boldsymbol{I} - \boldsymbol{H}(\lambda)) &= -(\mathrm{d}_\lambda \boldsymbol{H}(\lambda))^\top \widehat{\boldsymbol{W}}(\boldsymbol{I} - \boldsymbol{H}(\lambda)) - (\boldsymbol{I} - \boldsymbol{H}(\lambda))^\top \widehat{\boldsymbol{W}}\,\mathrm{d}_\lambda \boldsymbol{H}(\lambda) \\ &= (\widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}}(\boldsymbol{I} - \boldsymbol{H}(\lambda)) + (\boldsymbol{I} - \boldsymbol{H}(\lambda))^\top \widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}})\,\mathrm{d}\lambda,\end{aligned} \tag{B.19}$$

therefore

$$\mathrm{d}_\lambda(1 - \mathsf{edf}(\lambda)/n)^{-2} = \frac{2/n}{(1 - \mathsf{edf}(\lambda)/n)^3}\,\mathrm{tr}(\mathrm{d}_\lambda \boldsymbol{H}(\lambda)) = -\frac{2/n}{(1 - \mathsf{edf}(\lambda)/n)^3}\,\mathrm{tr}(\boldsymbol{G}\widehat{\boldsymbol{W}})\,\mathrm{d}\lambda. \tag{B.20}$$

Substituting equations (B.19) and (B.20) into (B.18) yields

$$\begin{aligned}\mathrm{d}^2_{\lambda\omega} V(\lambda, \boldsymbol{\omega}) = {} & \frac{2/n}{(1 - \mathsf{edf}(\lambda)/n)^2}\Big\{\boldsymbol{Y}^\top(\boldsymbol{\omega})(\widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}}(\boldsymbol{I} - \boldsymbol{H}(\lambda)) + (\boldsymbol{I} - \boldsymbol{H}(\lambda))^\top \widehat{\boldsymbol{W}}\boldsymbol{G}\widehat{\boldsymbol{W}}) \\ & - \frac{2/n}{(1 - \mathsf{edf}(\lambda)/n)}\,\mathrm{tr}(\boldsymbol{G}\widehat{\boldsymbol{W}})\boldsymbol{Y}^\top(\boldsymbol{\omega})(\boldsymbol{I} - \boldsymbol{H}(\lambda))^\top \widehat{\boldsymbol{W}}(\boldsymbol{I} - \boldsymbol{H}(\lambda))\Big\}\,\mathrm{d}\lambda\,\mathrm{d}\boldsymbol{\omega}.\end{aligned} \tag{B.21}$$

Evaluating (B.21) at $\lambda = \widehat{\lambda}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and applying the second identification theorem by Magnus and Neudecker (1999), we obtain $\partial^2 V(\lambda, \boldsymbol{\omega}_0)/\partial\lambda\partial\boldsymbol{\omega}^\top|_{\lambda=\widehat{\lambda}}$ and the Proposition 5 is verified. □

Appendix C. Real data example

We consider the balloon dataset reported by Davies and Gather (1993) and previously analyzed by Kovac and Silverman (2000), Lee and Oh (2007) and Tharmaratnam et al. (2010). The data are radiation measurements from the sun, taken during the flight of a weather balloon. Due to the rotation of the balloon, or for some other reasons, a large amount of outliers were introduced because the measuring instrument was occasionally blocked from the sun. The dataset is included in the R package `ftnonpar` available from CRAN repository. The sample size equals 4984. Following the model settings from Tharmaratnam et al. (2010), we applied P-splines using B-splines of third-degree and second-order of penalty, with 35 knots spread equally. The model was fitted using the routine heavyPS from the R package `heavy`, also avaliable at CRAN. All the computations were done on an IBM x3650 M4 Server with 2 Intel Xeon E5-2670 processors and 128 GB of RAM Quadro 6000.

Table 1. Estimation summary for the balloon dataset under three fitted models.

Model	$\widehat{\nu}$	$\widehat{\lambda}$	$\ell_\lambda(\widehat{\boldsymbol{\theta}}; \boldsymbol{Y}_{\text{obs}})$	edf	WGCV	iterations	time (sec.)
Normal	—	0.3284	2470.705	27.1735	0.0220	3	0.060
Slash	0.4306	0.0081	7503.993	30.2584	0.0001	101	3.529
Student-t	0.9712	0.0230	7513.882	30.3645	0.0003	84	3.799

As expected, the estimation procedure based on heavy-tailed distributions produces a curve estimate that is not affected by the outlying observations (see Figure 1 and Table 1). On the other hand, the curve estimate obtained under the normality assumption suffers from the presence of the outliers. That is, the whole estimated curve was pulled downwardly away from the majority of the observations. Another interesting feature of the proposed estimation procedure is that the implementation is fairly simple and ensure a reasonable computational speed (Compare with the numerical experiments reported by Staudenmayer et al., 2009).

We conducted an additional experiment, considering a sequence of values for the number of knots. Figures 2 to 7 show the fitted curves for the balloon dataset considering 15, 20, 25, 30, 35 and 40 knots. Table 2 presents the estimation results for this dataset under normal, slash and Student-t errors. It is interesting to note that for any number of knots the fitted curve under gaussian errors is strongly affected by the outlying observations, however the fitted curve obtained using heavy-tailed distributions do not suffer from this phenomenon. An unexpected finding is that the number of knots have an extremely large influence on the fitted curve under the normality assumption, but this influence is reduced when we consider distributions with heavier tails than the normal ones

Appendix D. Simulation study

A small simulation study was conducted to illustrate the practical performance of the proposed estimation procedure under a severe contamination scheme. The design of the simulation study is based on the used by Tharmaratnam et al. (2010) (see also Cantoni and Ronchetti, 2001; Lee and Oh, 2007). We have used the following test function

$$Y_i = \sin(2\pi(1 - x_i)^2) + \epsilon_i, \qquad i = 1, \ldots, n.$$

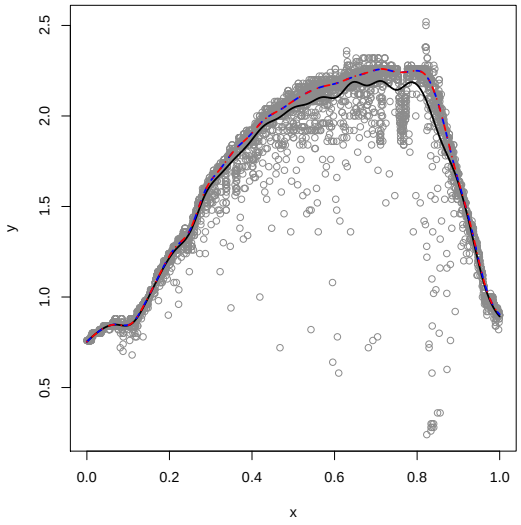

FIGURE 1. Fitted curve for the balloon dataset under three distributional assumptions: (—) normal, (- · -) Student-t and (- -) slash models.

TABLE 2. Summary for the balloon dataset under three fitted models, considering different number of knots.

Number	normal		slash			Student-t		
of knots	$\widehat{\lambda}$	$\ell_\lambda(\widehat{\boldsymbol{\theta}};\boldsymbol{Y}_{\text{obs}})$	$\widehat{\lambda}$	$\widehat{\nu}$	$\ell_\lambda(\widehat{\boldsymbol{\theta}};\boldsymbol{Y}_{\text{obs}})$	$\widehat{\lambda}$	$\widehat{\nu}$	$\ell_\lambda(\widehat{\boldsymbol{\theta}};\boldsymbol{Y}_{\text{obs}})$
15	0.010	2408.03	0.000	0.499	6721.26	0.000	1.119	6742.48
20	0.011	2419.76	0.000	0.481	6945.77	0.001	1.076	6965.06
25	0.018	2439.09	0.000	0.432	7349.22	0.000	0.966	7375.85
30	0.657	2427.79	0.000	0.437	7382.80	0.001	0.976	7406.13
35	0.328	2470.71	0.008	0.431	7504.00	0.023	0.971	7513.88
40	0.089	2515.78	0.001	0.434	7554.08	0.001	0.967	7573.12

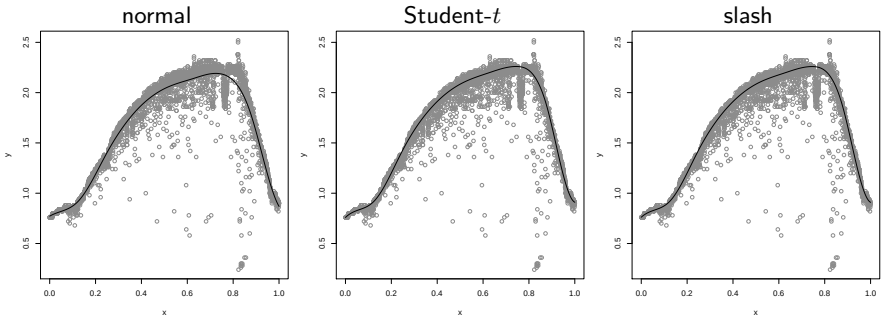

FIGURE 2. Fitted curve for the normal, Student-t and slash models using 15 knots.

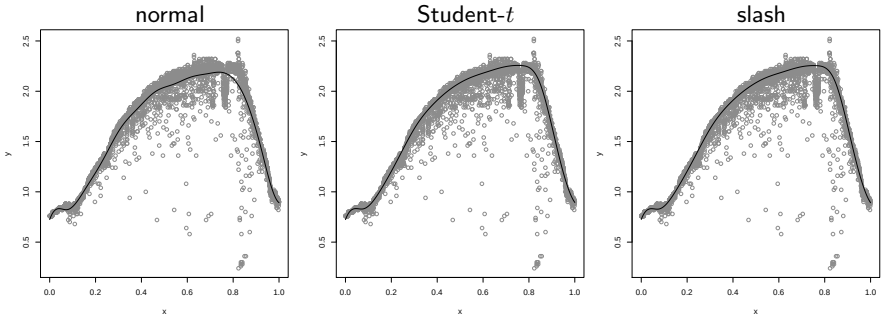

FIGURE 3. Fitted curve for the normal, Student-t and slash models using 20 knots.

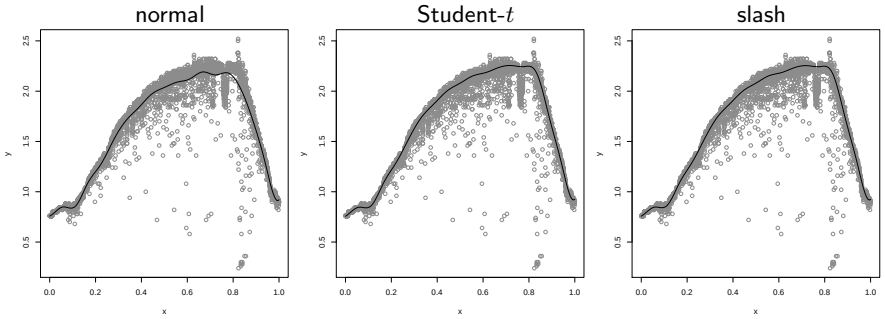

FIGURE 4. Fitted curve for the normal, Student-t and slash models using 25 knots.

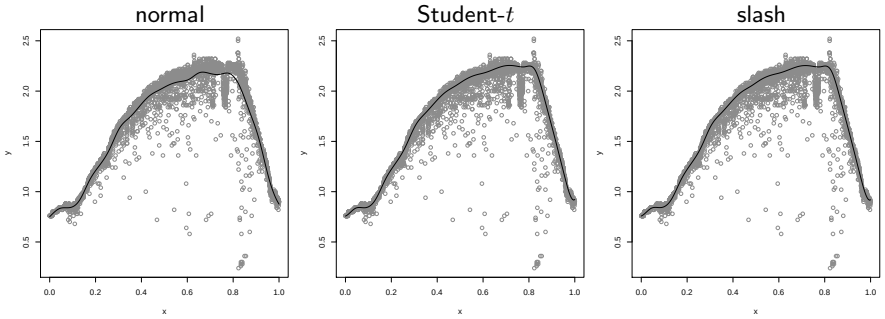

FIGURE 5. Fitted curve for the normal, Student-t and slash models using 30 knots.

We generated $M = 1000$ datasets of a sample size $n = 100$ from the model above with design points $x_1, \ldots, x_n$ generated independently from the uniform distribution $\mathcal{U}(-1, 1)$, the random disturbances $\{\epsilon_i\}$ were generated from the contaminated normal distribution

$$(1 - \delta)\mathcal{N}(0, 0.7^2) + \delta\mathcal{N}(20, 2^2),$$

for $\delta = 0\%, 5\%, 10\%, 25\%$ and 40% of outliers. We applied P-splines considering B-splines of third-degree and second-order of penalty, with 25 knots spread equally according to the quantiles of the data. The normal, slash and Student-t distributions were considered, fixing the degrees of freedom for the slash and Student-t

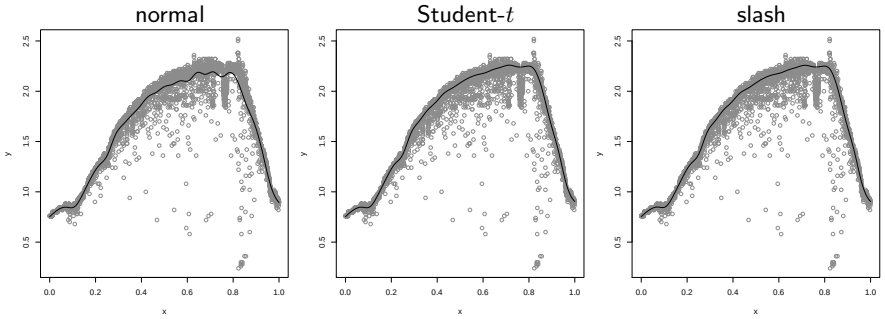

FIGURE 6. Fitted curve for the normal, Student-t and slash models using 35 knots.

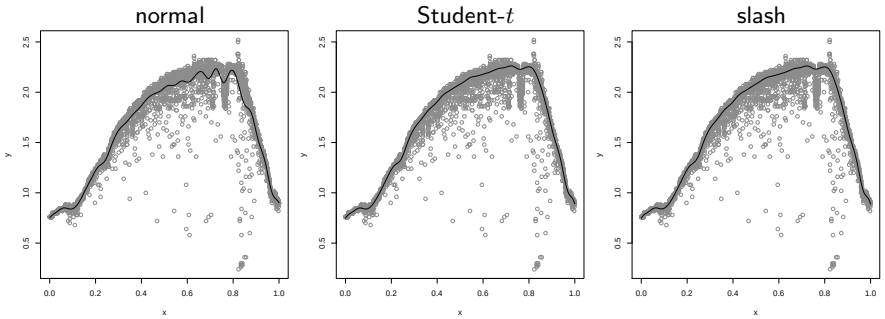

FIGURE 7. Fitted curve for the normal, Student-t and slash models using 40 knots.

at $1/2, 1$ and 2, and $1, 2$, and 4, respectively. The smoothing parameter is chosen by minimizing the weighted generalized cross-validation (WGCV) criterion defined in Section 2.2. Figure 8 displays two typical datasets together with the fitted curve from the gaussian and Student-t models (fitted curve for the slash model is similar to the obtained under Student-t errors and is not shown here), unlike the P-splines considering the assumption of normality the penalized spline estimator under Student-t errors with 1 degrees of freedom remains close to the true regression function (solid line) for both situations, without and in presence of outliers.

To assess the performance of the proposed procedure we compute the mean squared error (MSE) for each simulated dataset j ($j = 1, \ldots, M$), as

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

with $M = 1000$. We ran our simulation experiment on an IBM x3650 M4 Server with 2 Intel Xeon E5-2670 processors and 128 GB of RAM Quadro 6000. The total computation time was 16 hours, 56 minutes and 5.13 seconds.

Figures 9 to 11 present the boxplots of MSE in the logarithmic scale for several contamination percentages considering the P-splines estimation under normal, Student-t and slash distributions. From these plots it is clear that when we are in presence of outliers, the proposed procedure produces better estimates. In fact our findings are very similar to those reported by Tharmaratnam et al. (2010). However, it should be noted that the penalized spline estimator under heavy-tailed

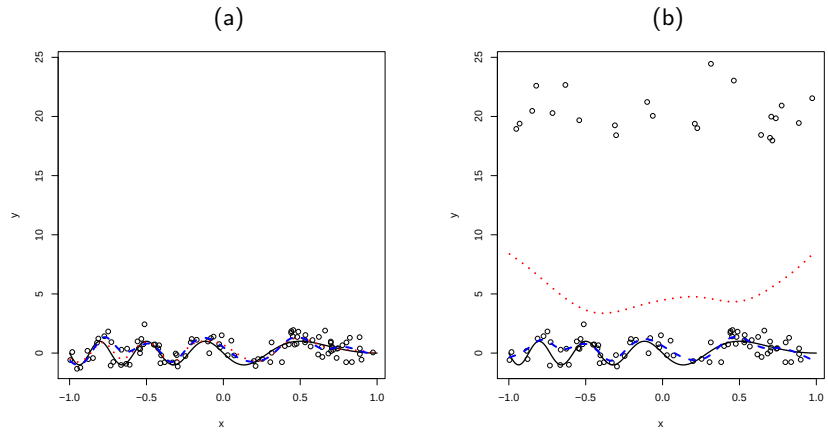

FIGURE 8. Fitted curve for the Cantoni and Ronchetti (2001) test function (a) without outliers and (b) with 25% of outliers from $\mathcal{N}(20, 2^2)$. True function $\sin(2\pi(1-x)^2)$ (solid line), fitted curve using normal errors (penalized LS-estimation) (dotted) and assuming Student-t model with $\nu = 1$ (dashed).

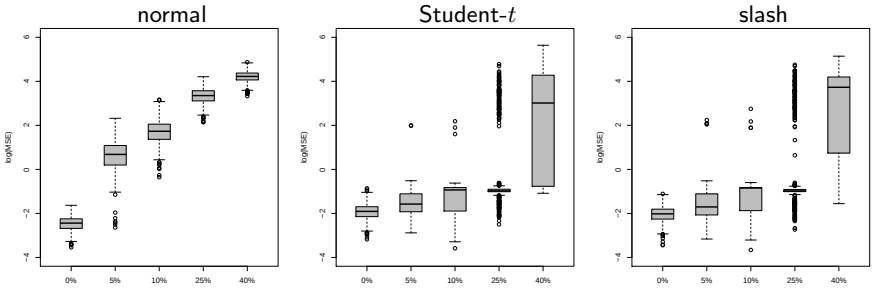

FIGURE 9. Boxplot of MSE using normal, Student-t ($\nu = 1$) and slash ($\nu = \frac{1}{2}$) models considering several contamination levels for the Cantoni and Ronchetti (2001) test function.

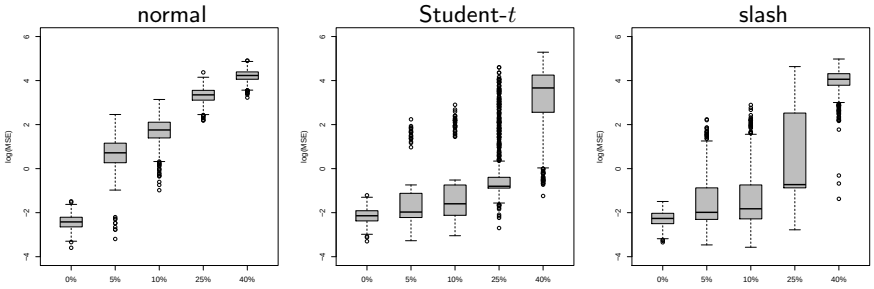

FIGURE 10. Boxplot of MSE using normal, Student-t ($\nu = 2$) and slash ($\nu = 1$) models considering several contamination levels for the Cantoni and Ronchetti (2001) test function.

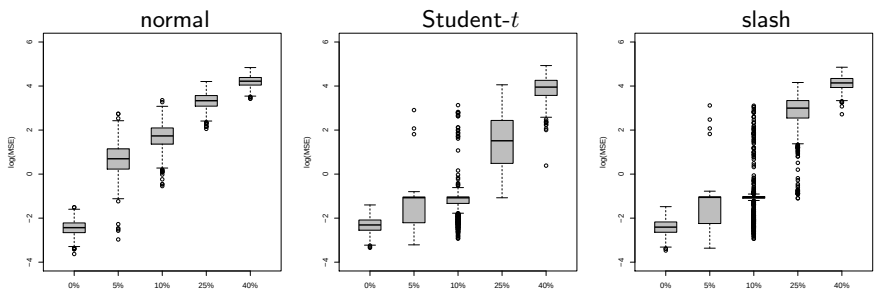

FIGURE 11. Boxplot of MSE using normal, Student-t ($\nu = 4$) and slash ($\nu = 2$) models considering several contamination levels for the Cantoni and Ronchetti (2001) test function.

distributions remains stable as the proportion of contamination increases only for small (and fixed) degrees of freedom. In our simulation experiment the penalized spline estimator's MSEs grows rapidly after 10% of outliers for degrees of freedom as small as 4 or 2, for the Student-t or slash distributions, respectively. In fact, as is discussed by Lange et al. (1989) modeling with distributions with heavier tails than the normal one is not the panacea for all robustness problems. In particular we can expect difficulties in P-splines under scale mixtures of normal distributions for data with extreme outliers.

REFERENCES

BRENT, R.P. (1973). *Algorithms for Minimization Without Derivatives*. Dover, New York.

CANTONI, E., RONCHETTI, E. (2001). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing* **11**, 141-146.

DAVIES, L., GATHER, U. (1993). The identification of multiple outliers (with discussion). *Journal of the American Statistical Association* **88**, 782-801.

DONGARRA, J.J., BUNCH, J.R., MOLER, C.B., STEWART, G.W. (1979). *Linpack Users Guide*. SIAM Publications, Philadelphia.

KOVAC, A., SILVERMAN, B.W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association* **95**, 172-183.

LANGE, K., LITTLE, R.J.A., TAYLOR, J.M.G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881-896.

LAWSON, C.L., HANSON, R.J., KINCAID, D., KROGH, F.T. (1979). Basic linear algebra subprograms for FORTRAN usage. *ACM Transactions on Mathematical Software* **5**, 308-323.

LEE, T.C.M., OH, H.S. (2007) Robust penalized regression spline fitting with application to additive mixed modeling. *Computational Statistics* **22**, 159-171.

MAGNUS, J.R., NEUDECKER, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, Chichester.

R CORE TEAM (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 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.

STAUDENMAYER, J., LAKE, E.E., 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., Salibián-Barrera, M. (2010). S-estimation for penalized regression splines. *Journal of Computational and Graphical Statistics* **19**, 609-625.

Departamento de Matemática, Universidad Técnica Federico Santa María, Chile
Current address: Departamento de Matemática, Universidad Técnica Federico Santa María, Av. España 1680, Casilla 110-V, Valparaíso, Chile
E-mail address: felipe.osorios@usm.cl