Computational Statistics
https://doi.org/10.1007/s00180-021-01106-2

ORIGINAL PAPER

Additive models with autoregressive symmetric errors based on penalized regression splines

Rodrigo A. Oliveira[1] · **Gilberto A. Paula**[2]

Received: 19 March 2020 / Accepted: 19 April 2021

Abstract

In this paper additive models with p-order autoregressive conditional symmetric errors based on penalized regression splines are proposed for modeling trend and seasonality in time series. The aim with this kind of approach is try to model the autocorrelation and seasonality properly to assess the existence of a significant trend. A backfitting iterative process jointly with a quasi-Newton algorithm are developed for estimating the additive components, the dispersion parameter and the autocorrelation coefficients. The effective degrees of freedom concerning the fitting are derived from an appropriate smoother. Inferential results and selection model procedures are proposed as well as some diagnostic methods, such as residual analysis based on the conditional quantile residual and sensitivity studies based on the local influence approach. Simulations studies are performed to assess the large sample behavior of the maximum penalized likelihood estimators. Finally, the methodology is applied for modeling the daily average temperature of San Francisco city from January 1995 to April 2020.

Keywords Cubic splines · Cyclic splines · Daily temperature · Model checking · Penalized likelihood · Student-t models · Robust estimation

Mathematics Subject Classification 62G08 · 62J05 · 62J20

✉ Gilberto A. Paula
giapaula@ime.usp.br

Rodrigo A. Oliveira
rodrigo.sef@hotmail.com

[1] Hospital das Clínicas, Empresa Brasileira de Serviços Hospitalares, Universidade Federal de Goiás, Goiânia, Goiás, Brazil

[2] Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil

Published online: 03 May 2021

1 Introduction

This paper proposes additive models with p-order autoregressive (AR(p)) conditional symmetric errors based on penalized regression splines for modeling trend and seasonality in time series by cubic and cyclic regression splines, respectively. The aim with this kind of approach is try to model the autocorrelation and seasonality properly to assess the existence of a significant trend. The assumption of AR(p) conditional symmetric errors allows to study a wide variety of time series by considering shorter- and heavier-tailed error distributions than the normal one. Some authors have discussed estimation and diagnostic in parametric regression models with autoregressive symmetric errors. For instance, Liu (2004) derived some procedures in conditional heteroscedastic time series models under elliptical errors, whereas Paula et al. (2009) gave emphasis on diagnostic methods in linear models with AR(1) elliptical errors and Cao et al. (2010) on the assessment of heteroscedasticity and autocorrelation in nonlinear models with AR(1) symmetric errors. Under the context of time series semiparametric models with autocorrelated errors, Relvas and Paula (2016) derived a Fisher scoring iterative procedure as well as some diagnostic procedures in partially linear models with AR(1) conditional symmetric errors, Liu et al. (2010) studied nonparametric transfer function models with ARMA errors and Huang et al. (2016, 2019) considered ARMA errors for semiparametric time series modeling.

The proposed model is defined in Sect. 2 and a penalized log-likelihood function is considered in Sect. 3 for the parameter estimation. The score function and the Fisher information matrix for the parameters of interest are derived in Sect. 4 as well as a backfitting (Gauss-Seidel) iterative process jointly with a quasi-Newton algorithm for obtaining the maximum penalized likelihood estimates (MPLEs). Inferential procedures are discussed in Sect. 5 and the effective degrees of freedom are derived from an appropriate smoother in Sect. 6. Usual methods for model selection are described in Sect. 7 and some diagnostic procedures, such as residual analysis based on the conditional quantile residual and sensitivity studies based on the local influence approach, are derived in Sect. 8. Simulation studies for assessing the large sample behavior of the MLPEs are described in Sect. 9. The methodology developed through the paper is applied in Sect. 10 for modeling the daily average temperature of San Francisco city from January 1995 to April 2020. The last section deals with concluding remarks whereas some technical results are described in Appendices A–D.

2 The model

Based on Cleveland et al. (1990) and Wood (2017, Sec. 7.7.2) we propose for modeling trend and seasonality of a daily time series the following additive model with AR(p) conditional symmetric errors:

$$y_i = f_T(t_i) + f_S(s_i) + \epsilon_i, \quad (i = 1, \ldots, n), \tag{1}$$

$$\epsilon_i = \rho_1 \epsilon_{i-1} + \cdots + \rho_p \epsilon_{i-p} + e_i, \tag{2}$$

where y_i denotes the ith observed response, $f_T(t_i)$ and $f_S(s_i)$ are smooth functions, t_i and s_i denote, respectively, the ith time and the time of the year respective to the ith time, $\rho_1, \ldots, \rho_p$ are the autocorrelation coefficients whereas e_i are independent symmetric errors with zero mean and dispersion parameter ϕ, namely $e_i \overset{iid}{\sim} S(0, \phi)$, and $y_i|(y_{i-1}, \ldots, y_{i-p}) \overset{ind}{\sim} S(\mu_i, \phi)$ with $\mu_i = \mathrm{E}(y_i) = f_T(t_i) + f_S(s_i)$, for $i = 1, \ldots, n$. In addition, we will assume that the smoothing functions $f_T(t)$ and $f_S(s)$ are approximated by cubic and cyclic cubic regression splines with fixed knots at the values t_j^0 $(j = 1, \ldots, r_T)$ and s_l^0 $(l = 1, \ldots, r_S)$, respectively.

There are many equivalent bases that can be used to represent cubic and cyclic splines. In the following, a penalized cubic regression spline and its cyclic version developed by Wood (2017) will be considered. The assumptions are that the cubic regression spline must be continuous to second derivatives at the knots t_j^0 $(j = 1, \ldots, r_T)$ and must have zero second derivative at the first and the last knots. Consequently, it may be written in the linear form

$$f_T(t) = \sum_{j=1}^{r_T} n_{T_j}(t)\gamma_{T_j},$$

where $n_{T_j}(t)$ denote the basis functions and γ_{T_j} are unknown parameters, for $j = 1, \ldots, r_T$. A cyclic version may be derived from $f_T(t)$ by assuming that the function as well as its first and second derivatives should match at the first and last knots of each cycle. For example, if we are considering a cycle as being a year, one should have that the function $f_S(t)$ as well as its first and second derivatives should math at the first and last knots of the year, which leads to $\gamma_{T_1} = \gamma_{T_{r_T}}$ in the notation of $f_T(t)$. So, in this case, defining $\boldsymbol{\gamma}_S = (\gamma_{S_1}, \ldots, \gamma_{S_{r_S-1}})^\top$ the cyclic regression spline may be also written in a linear form

$$f_S(s) = \sum_{l=1}^{r_S-1} n_{S_l}(s)\gamma_{S_l},$$

where $n_{S_l}(t)$ denote the basis functions and γ_{S_l} are unknown parameters, for $l = 1, \ldots, r_S - 1$. The quantities $n_{T_j}(t)$ $(j = 1, \ldots, r_T)$ and $n_{S_l}(t)$ $(l = 1, \ldots, r_S - 1)$ may be derived from Wood (2017, Table 5.1).

Then, one may rewritten model (1)–(2) in the matrix form

$$\mathbf{y} = \mathbf{N}_T\boldsymbol{\gamma}_T + \mathbf{N}_S\boldsymbol{\gamma}_S + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{y}$ is an $(n \times 1)$ vector of observed responses, $\mathbf{N}_T$ is an $(n \times r_T)$ basis function matrix with rows $\mathbf{n}_T(t_i) = (n_{T_1}(t_i), \ldots, n_{T_{r_T}}(t_i))^\top$ and that the (i, j)th element equals the parametrization of the regression spline in terms of its values at the knots t_j^0, $\mathbf{N}_S$ is an $(n \times (r_S - 1))$ basis function matrix with rows $\mathbf{n}_S(s_i) = (n_{S_1}(s_i), \ldots, n_{S_{r_S-1}}(s_i))^\top$ and that the (i, l)th element equals the parametrization of the regression spline in terms of its values at the knots s_l^0, for $i = 1, \ldots, n$, $j = 1, \ldots, r_T$ and $l = 1, \ldots, r_S - 1$,

whereas $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is an $(n \times 1)$ error vector. Due to the cyclic structure one has the restriction $\mathbf{n}_S(s_1) = \mathbf{n}_S(s_n)$. The unknown parameters to be estimated are $\boldsymbol{\gamma}_T = (\gamma_{T_1}, \ldots, \gamma_{T_{r_T}})^\top$ and $\boldsymbol{\gamma}_S = (\gamma_{S_1}, \ldots, \gamma_{S_{r_S-1}})^\top$.

Therefore, one has in (3) the model matrices

$$\mathbf{N}_T = \begin{bmatrix} n_{T_1}(t_1) & \ldots & n_{T_{r_T}}(t_1) \\ n_{T_1}(t_2) & \ldots & n_{T_{r_T}}(t_2) \\ \vdots & \ddots & \vdots \\ n_{T_1}(t_n) & \ldots & n_{T_{r_T}}(t_n) \end{bmatrix} \quad \text{end} \quad \mathbf{N}_S = \begin{bmatrix} n_{S_1}(s_1) & \ldots & n_{S_{r_S-1}}(s_1) \\ n_{S_1}(s_2) & \ldots & n_{S_{r_S-1}}(s_2) \\ \vdots & \ddots & \vdots \\ n_{S_1}(s_{n-1}) & \ldots & n_{S_{r_S-1}}(s_{n-1}) \\ n_{S_1}(s_1) & \ldots & n_{S_{r_S-1}}(s_1) \end{bmatrix}.$$

The probability density function of e_i is given by

$$h_e(e_i) = \frac{1}{\sqrt{\phi}} g(\delta_i), \quad e_i \in \mathcal{R},$$

where $\delta_i = \phi^{-1} e_i^2$, with $g : \mathcal{R} \to [0, \infty)$ such that $\int_0^\infty u^{-\frac{1}{2}} g(u) du = 1$ being known as a density generator.

The name density generator is due the fact that if u is a symmetric distribution, it does not necessarily have a density. Thus, if the density of u exists, it must assume the form $g(u^2)$ for some non-negative function $g(\cdot)$ of a scalar variable, that satisfies the condition $\int_0^\infty g(u^2) du = \int_0^\infty u^{-\frac{1}{2}} g(u) du = 1$ (see, for instance, Fang et al. 1990, Section 2.2.3). In this case one has $\mathrm{E}(e_i) = 0$ and $\mathrm{Var}(e_i) = \xi\phi$, where $\xi > 0$ is a constant that may be obtained from the expected value of the radial variable or from the derivative of the characteristic function. Cysneiros and Paula (2005) provides additional insight into the expressions of ξ for some symmetric distributions. This class includes all symmetric continuous distributions, such as normal, Student-t, power exponential, slash, logistic-I, logistic-II among others. For example, for the Student-t distribution with ν degrees of freedom one has $\xi = \nu/(\nu - 2)(\nu > 2)$. Herein, it is assumed that ξ is known or fixed, or estimated separately.

3 Penalized log-likelihood function

Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}_T^\top, \boldsymbol{\gamma}_S^\top, \phi, \boldsymbol{\rho}^\top)^\top \in \boldsymbol{\Theta} \subseteq \mathcal{R}^q$, with $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)^\top$ and $q = r_T + r_S + p$ being the number of parameters to be estimated. The regular log-likelihood function associated with $\boldsymbol{\theta}$ is given by

$$\mathrm{L}(\boldsymbol{\theta}) = -\frac{n}{2} \log(\phi) + \sum_{i=1}^{n} \log\{g(\delta_i)\}. \tag{4}$$

The maximization of (4) without imposing restrictions over the nonparametric functions $f_T(t)$ and $f_S(s)$ may cause overfitting and non identification of parameters. To solve this problem one may incorporate a penalty function over each nonparametric

component. Similarly to Hastie and Tibshirani (1990) (see also Green and Silverman 1994) we will assume that $f_T(t)$ and $f_S(s)$ are continuous to second derivatives and integrable second derivative, so the penalized log-likelihood function is given by

$$\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathrm{L}(\boldsymbol{\theta}) - \frac{\lambda_T}{2}\int_{a_T}^{b_T}\left[f_T''(t)\right]^2 dt - \frac{\lambda_S}{2}\int_{a_S}^{b_S}\left[f_S''(s)\right]^2 ds, \tag{5}$$

where $a_T \leq t \leq b_T$, $a_S \leq s \leq b_S$ and $\lambda_T > 0$ and $\lambda_S > 0$ are the smoothing parameters that will be estimated separately. It can also be shown, for example, in Lancaster and Salkauskas (1986) and Wood (2017) that

$$\int_{a_T}^{b_T}\left[f_T''(t)\right]^2 dt = \boldsymbol{\gamma}_T^\top \mathbf{M}_T \boldsymbol{\gamma}_T \quad \text{and} \quad \int_{a_S}^{b_S}\left[f_S''(s)\right]^2 ds = \boldsymbol{\gamma}_S^\top \mathbf{M}_S \boldsymbol{\gamma}_S,$$

where $\mathbf{M}_T = \mathbf{D}_T^\top \mathbf{B}_T^{-1}\mathbf{D}_T$ and $\mathbf{M}_S = \mathbf{D}_S^\top \mathbf{B}_S^{-1}\mathbf{D}_S$ are non-negative definite matrices of dimensions $(r_T \times r_T)$ and $((r_S - 1) \times (r_S - 1))$, respectively, named penalty matrices. These matrices depend only on the knots and are defined in Table 5.1 of Wood (2017).

Thus, the penalized log-likelihood function reduces to

$$\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathrm{L}(\boldsymbol{\theta}) - \frac{\lambda_T}{2}\boldsymbol{\gamma}_T^\top \mathbf{M}_T \boldsymbol{\gamma}_T - \frac{\lambda_S}{2}\boldsymbol{\gamma}_S^\top \mathbf{M}_S \boldsymbol{\gamma}_S, \tag{6}$$

where $\boldsymbol{\lambda} = (\lambda_T, \lambda_S)^\top$ denotes the smoothing parameter vector. In the next section we will derive a backfitting (Gauss-Seidel) iterative process jointly with a quasi-Newton algorithm for maximizing (6) for fixed $\boldsymbol{\lambda}$.

4 Parameter estimation

4.1 Penalized score function and penalized Fisher information matrix

The penalized score function of $\boldsymbol{\theta}$ may be expressed as $\mathbf{U}_p^\theta = \partial \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})/\partial\boldsymbol{\theta} = \mathbf{U}_\theta - \mathbf{M}(\boldsymbol{\lambda})\boldsymbol{\theta}$, where $\mathbf{U}_\theta$ is the regular score function given by

$$\mathbf{U}_\theta = \begin{pmatrix} \mathbf{U}_{\gamma_T} \\ \mathbf{U}_{\gamma_S} \\ \mathrm{U}_\phi \\ \mathrm{U}_{\rho_1} \\ \vdots \\ \mathrm{U}_{\rho_p} \end{pmatrix} = \phi^{-1}\begin{pmatrix} (\mathbf{A}\mathbf{N}_T)^\top \mathbf{D}_v \mathbf{A}\boldsymbol{\epsilon} \\ (\mathbf{A}\mathbf{N}_S)^\top \mathbf{D}_v \mathbf{A}\boldsymbol{\epsilon} \\ \frac{1}{2}\mathbf{1}_n^\top(\mathbf{D}_m \mathbf{1}_n - \mathbf{1}_n) \\ -(\mathbf{C}_1\boldsymbol{\epsilon})^\top \mathbf{D}_v \mathbf{A}\boldsymbol{\epsilon} \\ \vdots \\ -(\mathbf{C}_p\boldsymbol{\epsilon})^\top \mathbf{D}_v \mathbf{A}\boldsymbol{\epsilon} \end{pmatrix},$$

$\mathbf{M}(\boldsymbol{\lambda}) = \text{blockdiag}\left\{\lambda_T \mathbf{M}_T, \lambda_S \mathbf{M}_S, \mathbf{0}_{p+1}\right\}$ with $\mathbf{0}_{p+1}$ being a $(p+1 \times p+1)$ matrix

of zeros, $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{N}\boldsymbol{\gamma}$ with $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_T^\top, \boldsymbol{\gamma}_S^\top)^\top$ and $\mathbf{N} = (\mathbf{N}_T, \mathbf{N}_S)$. $\mathbf{A}$ and $\mathbf{C}$ are $(n \times n)$ matrices defined as

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \ldots & 0 & 0 \ldots & 0 & 0\,0 \\ -\rho_1 & 1 & 0 \ldots & 0 & 0 \ldots & 0 & 0\,0 \\ -\rho_2 & -\rho_1 & 1 \ldots & 0 & 0 \ldots & 0 & 0\,0 \\ \vdots & \vdots & \vdots \ddots & \vdots & \vdots\ \vdots & \vdots & \vdots\ \vdots \\ -\rho_p & -\rho_{(p-1)} & -\rho_{(p-2)} \ldots & -\rho_1 & 1 \cdots & 0 & 0\,0 \\ 0 & -\rho_p & -\rho_{(p-1)} \cdots & -\rho_2 & -\rho_1\ 1 & \cdots & 0\,0 \\ \vdots & \vdots & \vdots \ddots & \vdots & \vdots\ \vdots & \vdots & \vdots\ \vdots \\ 0 & \cdots & 0 \cdots & 0 & 0 \cdots & -\rho_2 & -\rho_1\ 1 \end{pmatrix}$$

and

$$\mathbf{C}_j = \begin{pmatrix} & & \mathbf{0}_{(j\times n)} & & \\ -1 & 0\,0 & \ldots & 0\,0 \\ 0 & -1\,0 & \ldots & 0\,0 \\ \vdots & \vdots\ \vdots & \ddots & \vdots\ \vdots \\ 0 & 0\,0 & \ldots & -1\,0 \end{pmatrix},$$

where $\mathbf{D}_v = \text{diag}\{v_1, \ldots, v_n\}$ with $v_i = -2W_g(\delta_i)$ named weights and $\mathbf{D}_m = \text{diag}\{m_1, \ldots, m_n\}$ with $m_i = v_i\delta_i$, for $i = 1, \ldots, n$, whereas $W_g(\delta) = g'(\delta)/g(\delta)$ and $\mathbf{1}_n$ is an $(n \times 1)$ vector of ones (see details in Appendix A). Cysneiros and Paula (2005) provide a table with the quantities $W_g(\delta)$ and $W_g'(\delta)$ for some symmetric distributions.

In addition, the penalized Fisher information matrix of $\boldsymbol{\theta}$ is defined as $\mathbf{K}_p^{\theta\theta} = -\text{E}\left\{\partial \text{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top\right\} = \mathbf{K}_{\theta\theta} - \mathbf{M}(\boldsymbol{\lambda})$, where

$$\mathbf{K}_{\theta\theta} = \begin{pmatrix} \mathbf{K}_{\gamma_T\gamma_T} & \mathbf{K}_{\gamma_T\gamma_S} & \mathbf{0} & \mathbf{0} \\ \mathbf{K}_{\gamma_S\gamma_T} & \mathbf{K}_{\gamma_S\gamma_S} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \text{K}_{\phi\phi} & 0 \\ \mathbf{0} & \mathbf{0} & 0 & \mathbf{K}_{\rho\rho} \end{pmatrix} = \text{blockdiag}\{\mathbf{K}_{\gamma\gamma}, \text{K}_{\phi\phi}, \mathbf{K}_{\rho\rho}\}$$

is the regular Fisher information matrix,

$$\begin{aligned} \mathbf{K}_{\gamma\gamma} &= \frac{4d_g}{\phi}\left(\mathbf{N}_A^\top\mathbf{N}_A\right), \\ \text{K}_{\phi\phi} &= \frac{n}{4\phi^2}(4f_g - 1) \text{ and} \\ \mathbf{K}_{\rho\rho} &= \frac{4d_g}{\phi}\boldsymbol{\Upsilon}_\rho \end{aligned}$$

with $\mathbf{N}_A = (\mathbf{A}\mathbf{N}_T, \mathbf{A}\mathbf{N}_S)$, $d_g = \text{E}\{W_g^2(z^2)z^2\}$, $f_g = \text{E}\{W_g^2(z^2)z^4\}$, $z \sim S(0, 1)$ and

$$\boldsymbol{\Upsilon}_{\rho} = \begin{pmatrix} \sum_{i=1}^{n-1} \mathrm{E}(\epsilon_i^2) & \sum_{i=1}^{n-1} \mathrm{E}(\epsilon_i \epsilon_{i+1}) & \dots & \sum_{i=1}^{n-p+1} \mathrm{E}(\epsilon_i \epsilon_{i+p-1}) \\ \sum_{i=1}^{n-1} \mathrm{E}(\epsilon_i \epsilon_{i+1}) & \sum_{i=1}^{n-2} \mathrm{E}(\epsilon_i^2) & \dots & \sum_{i=1}^{n-p+2} \mathrm{E}(\epsilon_i \epsilon_{i+p-2}) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n-p+1} \mathrm{E}(\epsilon_i \epsilon_{i+p-1}) & \sum_{i=1}^{n-p+2} \mathrm{E}(\epsilon_i \epsilon_{i+p-2}) & \dots & \sum_{i=1}^{n-p} \mathrm{E}(\epsilon_i^2) \end{pmatrix}$$

is a $(p \times p)$ matrix discussed in Appendix A. Note that one has orthogonality between $\boldsymbol{\gamma}$ and $(\phi, \boldsymbol{\rho}^\top)^\top$, see details in Appendices A and D.

4.2 Iterative process

Similarly to Hastie and Tibshirani (1990, Ch.5) and Wood (2017, Ch.6) we will derive a backfitting (Gauss-Seidel) iterative process for obtaining the MPLE $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}_T^\top, \widehat{\boldsymbol{\gamma}}_S^\top)^\top$ alternated with a quasi-Newton algorithm for obtaining the MPLEs of ϕ and $\boldsymbol{\rho}$. Assuming $\boldsymbol{\lambda}$ fixed one has to solve the following equations:

$$\left\{(\mathbf{A}\mathbf{N}_T)^\top \mathbf{D}_v (\mathbf{A}\mathbf{N}_T) + \phi\lambda_T \mathbf{M}_T\right\} \boldsymbol{\gamma}_T = (\mathbf{A}\mathbf{N}_T)^\top \mathbf{D}_v \mathbf{A}(\mathbf{y} - \mathbf{N}_S \boldsymbol{\gamma}_S)$$

and

$$\left\{(\mathbf{A}\mathbf{N}_S)^\top \mathbf{D}_v (\mathbf{A}\mathbf{N}_S) + \phi\lambda_S \mathbf{M}_S\right\} \boldsymbol{\gamma}_S = (\mathbf{A}\mathbf{N}_S)^\top \mathbf{D}_v \mathbf{A}(\mathbf{y} - \mathbf{N}_T \boldsymbol{\gamma}_T),$$

which may be simplified to

$$\boldsymbol{\gamma}_T = \mathbf{S}_T(\lambda_T)\mathbf{A}\left\{\mathbf{y} - \mathbf{N}_S \boldsymbol{\gamma}_S\right\} \text{ and } \boldsymbol{\gamma}_S = \mathbf{S}_S(\lambda_S)\mathbf{A}\left\{\mathbf{y} - \mathbf{N}_T \boldsymbol{\gamma}_T\right\},$$

where $\mathbf{S}_T(\lambda_T) = \left\{(\mathbf{A}\mathbf{N}_T)^\top \mathbf{D}_v (\mathbf{A}\mathbf{N}_T) + \phi\lambda_T \mathbf{M}_T\right\}^{-1} (\mathbf{A}\mathbf{N}_T)^\top \mathbf{D}_v$ and $\mathbf{S}_S(\lambda_S) = \left\{(\mathbf{A}\mathbf{N}_S)^\top \mathbf{D}_v (\mathbf{A}\mathbf{N}_S) + \phi\lambda_S \mathbf{M}_S\right\}^{-1} (\mathbf{A}\mathbf{N}_S)^\top \mathbf{D}_v$.

Then, we will propose, for fixed ϕ, $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$, the following iterative process for obtaining the MLPE $\widehat{\boldsymbol{\gamma}}$:

Step 1: Define the knots, and then compute $\mathbf{N}_T$, $\mathbf{N}_S$, $\mathbf{M}_T$ and $\mathbf{M}_S$.
Step 2: Initialize a counter as $u = 0$, set the initial value at $\boldsymbol{\theta}^{(0)}$.
Step 3: Based on $\boldsymbol{\theta}^{(u)}$ do the following:

(a) From the current value $\boldsymbol{\theta}^{(0)} = \left({\boldsymbol{\gamma}_T}^{(0)^\top}, {\boldsymbol{\gamma}_S}^{(0)^\top}, \phi^{(0)}, {\rho}^{(0)^\top}\right)^\top$ obtaining $\mathbf{A}$, the weights $v_i^{(0)} = v_i \mid_{\theta^{(0)}}$ and $\mathbf{D}_v^{(0)} = \mathrm{diag}_{1 \le i \le n}\left\{v_i^{(0)}\right\}$. Then, calculate

$$\mathbf{S}_T^{(0)}(\lambda_T) = \left\{(\mathbf{A}\mathbf{N}_T)^\top \mathbf{D}_v^{(0)} (\mathbf{A}\mathbf{N}_T) + \phi\lambda_T \mathbf{M}_T\right\}^{-1} (\mathbf{A}\mathbf{N}_T)^\top \mathbf{D}_v^{(0)}$$

and

$$\mathbf{S}_S^{(0)}(\lambda_S) = \left\{(\mathbf{A}\mathbf{N}_S)^\top \mathbf{D}_v^{(0)} (\mathbf{A}\mathbf{N}_S) + \phi\lambda_S \mathbf{M}_S\right\}^{-1} (\mathbf{A}\mathbf{N}_S)^\top \mathbf{D}_v^{(0)}.$$

(b) Compute the following expressions for $\boldsymbol{\gamma}_T^{(u+1)}$ and $\boldsymbol{\gamma}_S^{(u+1)}$:

$$\boldsymbol{\gamma}_T^{(u+1)} = \mathbf{S}_T^{(u)}(\lambda_T)\mathbf{A}\left\{\mathbf{y} - \mathbf{N}_S\boldsymbol{\gamma}_S^{(u)}\right\}$$

and

$$\boldsymbol{\gamma}_S^{(u+1)} = \mathbf{S}_S^{(u)}(\lambda_S)\mathbf{A}\left\{\mathbf{y} - \mathbf{N}_T\boldsymbol{\gamma}_T^{(u+1)}\right\},$$

for $u = 0, 1, \ldots$. Update $\mathbf{S}_T^{(u)}(\lambda_T)$ and $\mathbf{S}_S^{(u)}(\lambda_S)$. Repeat (b) replacing $\boldsymbol{\gamma}^{(u)}$ by $\boldsymbol{\gamma}^{(u+1)}$, respectively, until the convergence for any criterion.

Step 4: Iterating the step 3(b) with the iterative process below that applies the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method (see Davidon 1991, and Mittelhammer et al. 2000, p.199) or the L-BFGS-B method (Byrd et al. 1995) for obtaining the MPLE of $\zeta = (\phi, \boldsymbol{\rho}^\top)^\top$:

$$\zeta^{(s+1)} = \arg\max_{(\phi,\boldsymbol{\rho})} \mathrm{L}_p(\widehat{\boldsymbol{\gamma}}^{(u+1)}, \phi^{(s)}, \boldsymbol{\rho}^{(s)}), \quad s = 0, 1, 2, \ldots.$$

The corresponding standard errors may be approximated by the square root of the principal diagonal elements of the observed Fisher information matrix inverse (Efron and Hinkley 1978) or by evaluated $\mathbf{K}_{\rho\rho}$ and $\mathrm{K}_{\phi\phi}$ at the parameter estimates.

For the Student$-t$ distribution with ν degrees of freedom the current weight $v_i^{(u)} = (\nu+1)/(\nu+\delta_i^{(u)})$, with $\delta_i^{(u)} = \delta_i \mid_\theta^{(u)}$, is inversely proportional to the distance between the observed value y_i and its current predicted values $\mu_i^{(u)}$, so that outlying observations tend to have small weights in the estimation process. Similar interpretation may be applied for other heavier-tailed error distributions.

5 Inferential procedures

Linear models under normal and symmetric errors have similar regularity conditions for large sample to ensure consistency, efficiency and normality of the MPLEs. As for the parametric case, the approximate variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$ may be derived from the inverse of the expected information matrix. Essentially, $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}) = \left(\mathbf{K}_p^{\theta\theta} \mid_{\widehat{\theta}}\right)^{-1}$, where $\mathbf{K}_p^{\theta\theta}$ was defined in Sect. 4.1. This approach has support on the Bayesian approach for linear models, as described in Wood (2017, p.293). If $\mathbf{y} = \mathbf{N}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ follows an n-variate normal distribution and an improper prior is assumed for $\boldsymbol{\gamma}$, then the posterior distribution $\boldsymbol{\gamma}|\mathbf{y}$ follows a multivariate normal distribution of mean $\hat{\boldsymbol{\gamma}}$ and variance-covariance matrix that corresponds to the inverse of the respective penalized Fisher information matrix for $\boldsymbol{\gamma}$. So, credible intervals may be constructed for any quantities derived from $\boldsymbol{\gamma}$. This approach may be extended, for large n, for the symmetric class.

Considering the orthogonality between $\boldsymbol{\gamma}$ and $(\phi, \boldsymbol{\rho})^\top$ as well as between ϕ and $\boldsymbol{\rho}$, one obtains

$$\widehat{\text{Var}}(\widehat{\boldsymbol{\gamma}}) = (\mathbf{K}_p^{\gamma\gamma} \mid_{\widehat{\theta}})^{-1},$$

where $\mathbf{K}_p^{\gamma\gamma} = \mathbf{K}_{\gamma\gamma} + \mathbf{M}_\gamma(\boldsymbol{\lambda})$ with $\mathbf{M}_\gamma(\boldsymbol{\lambda}) = \text{blockdiag}\{\lambda_T \mathbf{M}_T, \lambda_S \mathbf{M}_S\}$. Furthermore, $\widehat{\text{Var}}(\hat{\phi}) = \text{K}^{-1}_{\phi\phi} \mid_{\widehat{\theta}}$ and $\widehat{\text{Var}}(\hat{\boldsymbol{\rho}}) = \mathbf{K}^{-1}_{\rho\rho} \mid_{\widehat{\theta}}$. Since $\mathbf{f}_T = \mathbf{N}_T \boldsymbol{\gamma}_T$ and $\mathbf{f}_S = \mathbf{N}_S \boldsymbol{\gamma}_S$, where $\mathbf{f}_T = (f_T(t_1), \ldots, f_T(t_n))^\top$ and $\mathbf{f}_S = (f_S(s_1), \ldots, f_S(s_n))^\top$, one may derive the approximate variance-covariance matrices

$$\text{Var}(\widehat{\mathbf{f}}_T) = \mathbf{N}_T \text{Var}(\widehat{\boldsymbol{\gamma}}_T)\mathbf{N}_T^\top \;\text{ and }\; \text{Var}(\widehat{\mathbf{f}}_S) = \mathbf{N}_S \text{Var}(\widehat{\boldsymbol{\gamma}}_S)\mathbf{N}_S^\top.$$

Then, similarly to Vanegas and Paula (2016) pointwise confidence bands may be performed for $(f_T(t_1), \ldots, f_T(t_n))^\top$ and $(f_S(s_1), \ldots, f_S(s_n))^\top$.

6 Effective degrees of freedom

Since penalization of the smooth functions $f_T(t)$ and $f_S(s)$ leads to a shrinkage of the MPLEs $\widehat{\boldsymbol{\gamma}}_T$ and $\widehat{\boldsymbol{\gamma}}_S$ with respect to the MLEs, it is crucial for model selection and hypothesis testing the estimation of the total degrees of freedom corresponding to the MPLEs. Similarly to Hastie and Tibshirani (1990, Ch.5), Green and Silverman (1994, Ch.5) and more recently Wood (2017, Ch.5) we will derive below the total degrees of freedom corresponding to the MPLEs $\widehat{\boldsymbol{\gamma}}_T$ and $\widehat{\boldsymbol{\gamma}}_S$ from an appropriate smoother.

At the solution of the estimating equations $\mathbf{U}_p^\theta = \mathbf{0}$ one has that

$$\begin{aligned}\widehat{\boldsymbol{\gamma}} &= \left\{\widehat{\mathbf{N}}_A^\top \widehat{\mathbf{D}}_v \widehat{\mathbf{N}}_A + \hat{\phi}\mathbf{M}_\gamma(\boldsymbol{\lambda})\right\}^{-1} \left\{\widehat{\mathbf{N}}_A^\top \widehat{\mathbf{D}}_v \widehat{\mathbf{N}}_A \widehat{\boldsymbol{\gamma}} + \widehat{\mathbf{N}}_A^\top \widehat{\mathbf{D}}_v \widehat{\mathbf{A}}(\mathbf{y} - \mathbf{N}\widehat{\boldsymbol{\gamma}})\right\} \\ &= \left\{\widehat{\mathbf{N}}_A^\top \widehat{\mathbf{D}}_v \widehat{\mathbf{N}}_A + \hat{\phi}\mathbf{M}_\gamma(\boldsymbol{\lambda})\right\}^{-1} \widehat{\mathbf{N}}_A^\top \widehat{\mathbf{D}}_v^{\frac{1}{2}} \widehat{\mathbf{D}}_v^{\frac{1}{2}} \widehat{\mathbf{A}}\mathbf{y}.\end{aligned}$$

Then, it follows the relationship $\widehat{\mathbf{D}}_v^{\frac{1}{2}} \widehat{\mathbf{N}}_A \widehat{\boldsymbol{\gamma}} = \widehat{\mathbf{D}}_v^{\frac{1}{2}} \widehat{\mathbf{A}} \mathbf{N} \widehat{\boldsymbol{\gamma}} = \widehat{\mathbf{D}}_v^{\frac{1}{2}} \widehat{\mathbf{A}} \widehat{\boldsymbol{\mu}} = \widehat{\mathbf{H}}(\boldsymbol{\lambda}) \widehat{\mathbf{D}}_v^{\frac{1}{2}} \widehat{\mathbf{A}}\mathbf{y}$, where

$$\widehat{\mathbf{H}}(\boldsymbol{\lambda}) = \widehat{\mathbf{D}}_v^{\frac{1}{2}} \widehat{\mathbf{N}}_A \left\{\widehat{\mathbf{N}}_A^\top \widehat{\mathbf{D}}_v \widehat{\mathbf{N}}_A + \widehat{\phi}\mathbf{M}_\gamma(\boldsymbol{\lambda})\right\}^{-1} \widehat{\mathbf{N}}_A^\top \widehat{\mathbf{D}}_v^{\frac{1}{2}},$$

is named linear smoother for $\boldsymbol{\lambda}$ fixed, whereas $\mathbf{D}_v^{\frac{1}{2}}\mathbf{A}\mathbf{y}$ may be interpreted as a weighted transformed response. In particular, under normal errors and $\boldsymbol{\rho} = \mathbf{0}$, one obtains $\widehat{\mathbf{H}}(\boldsymbol{\lambda}) = \mathbf{N}\{\mathbf{N}^\top\mathbf{N} + \hat{\phi}\mathbf{M}_\gamma(\boldsymbol{\lambda})\}^{-1}\mathbf{N}^\top$.

The effective degrees of freedom of smoother is defined as the sum of the eigenvalues of $\widehat{\mathbf{H}}(\boldsymbol{\lambda})$, namely $\text{df}_s(\boldsymbol{\lambda}) = \text{tr}\{\widehat{\mathbf{H}}(\boldsymbol{\lambda})\}$. From Eilers and Marx (1996) one has that

$$\begin{aligned}
\mathrm{df}_s(\boldsymbol{\lambda}) &= \mathrm{tr}\left\{\widehat{\mathbf{D}}_v^{\frac{1}{2}}\widehat{\mathbf{N}}_A\left(\widehat{\mathbf{N}}_A^\top\widehat{\mathbf{D}}_v\widehat{\mathbf{N}}_A + \widehat{\phi}\mathbf{M}_\gamma(\boldsymbol{\lambda})\right)^{-1}\widehat{\mathbf{N}}_A^\top\widehat{\mathbf{D}}_v^{\frac{1}{2}}\right\} \\
&= \mathrm{tr}\left\{\left(\widehat{\mathbf{N}}_A^\top\widehat{\mathbf{D}}_v\widehat{\mathbf{N}}_A + \widehat{\phi}\mathbf{M}_\gamma(\boldsymbol{\lambda})\right)^{-1}\widehat{\mathbf{N}}_A^\top\widehat{\mathbf{D}}_v\widehat{\mathbf{N}}_A\right\} \\
&= \mathrm{tr}\left\{\left(\mathbf{I}_{r_T+r_S} + \mathbf{Q}^{-\frac{1}{2}}\widehat{\phi}\mathbf{M}_\gamma(\boldsymbol{\lambda})\mathbf{Q}^{-\frac{1}{2}}\right)^{-1}\right\} \\
&= \sum_{i=1}^{r_T+r_S-1}\frac{1}{1+\alpha_i(\boldsymbol{\lambda})},
\end{aligned}$$

where $\alpha_i(\boldsymbol{\lambda}) \geq 0$ are the eigenvalues of the non-negative definite matrix $\mathbf{Q}^{-\frac{1}{2}}\widehat{\phi}\mathbf{M}_\gamma(\boldsymbol{\lambda})\mathbf{Q}^{-\frac{1}{2}}$ and $\mathbf{Q}^{\frac{1}{2}}\mathbf{Q}^{\frac{1}{2}} = \widehat{\mathbf{N}}_A^\top\widehat{\mathbf{D}}_v\widehat{\mathbf{N}}_A$, for $i = 1, \ldots, r_T + r_S - 1$. The effective degrees of freedom $\mathrm{df}_s(\lambda_T)$ and $\mathrm{df}_s(\lambda_S)$ correspond to the sum of the first r_T and the last $r_S - 1$ eigenvalues, respectively, of the linear smooth $\widehat{\mathbf{H}}(\boldsymbol{\lambda})$. Therefore the effective degrees of freedom is given by

$$\mathrm{df}(\boldsymbol{\lambda}) = \mathrm{df}_s(\boldsymbol{\lambda}) + p + 1.$$

7 Model selection

The Akaike information criterion (AIC) (Akaike 1973) or the Bayesian information criterion (BIC) (Schwarz 1978) may be used for selecting an appropriate symmetric error distribution in model (1)–(2) as well as the smoothing parameters λ_T and λ_S in the iterative process described in Sect. 4.2. Both criteria consist in minimizing the functions

$$\mathrm{AIC}(\boldsymbol{\lambda}) = -2\mathrm{L}_p(\widehat{\boldsymbol{\theta}}, \boldsymbol{\lambda}) + 2\mathrm{df}(\boldsymbol{\lambda}) \;\text{ and }\; \mathrm{BIC}(\boldsymbol{\lambda}) = -2\mathrm{L}_p(\widehat{\boldsymbol{\theta}}, \boldsymbol{\lambda}) + \log(n)\mathrm{df}(\boldsymbol{\lambda}).$$

Alternatively, the generalized cross-validation method (see, for instance, Wood (2017)) may be applied for selecting appropriate smoothing parameters. Such criteria consists in minimizing the function

$$\mathrm{GCV}(\boldsymbol{\lambda}) = \frac{n||\sqrt{\widehat{\mathbf{D}}_v}\widehat{\mathbf{A}}(\mathbf{y} - \widehat{\boldsymbol{\mu}})||^2}{\left[n - \mathrm{tr}\left\{\widehat{\mathbf{H}}(\boldsymbol{\lambda})\right\}\right]^2}.$$

The vector $\boldsymbol{\lambda}$ will be obtained by minimizing jointly $\mathrm{GCV}(\boldsymbol{\lambda})$ and $\mathrm{AIC}(\boldsymbol{\lambda})$ for a grid of the smoothing parameter values.

8 Diagnostic methods

8.1 Residuals analysis

Residual analysis is an important tool for detecting outlying observations and departures from the assumptions made for the error distribution in regression models. Among the various residuals proposed, the quantile residual (Dunn and Smyth 1996) may be easily derived when the cumulative distribution function (cdf) is known. Such residual for independent observations $(y_1, \ldots, y_n)^\top$ is defined as $r_{q_i} = \Phi^{-1}\{F_y(y_i; \widehat{\boldsymbol{\theta}})\}$, where $F_y(y_i; \boldsymbol{\theta})$ and $\Phi(\cdot)$ denote, respectively, the cdf of y_i and the cdf of N(0, 1), for $i = 1, \ldots, n$. For large n and under the postulated model $r_{q_1}, \ldots, r_{q_n}$ follow independent standard normal distributions.

However, according to Barros and Paula (2019), for correlated observations one should consider the independent conditional distributions with the respective conditional cdfs $F_{y_1}(y_1; \boldsymbol{\theta})$, $F_{y_2|y_1}(y_2; \boldsymbol{\theta}), \ldots, F_{y_n|(y_1,\ldots,y_{n-1})}(y_n; \boldsymbol{\theta})$. Then, the conditional quantile residuals become given by

$$r^c_{q_1} = \Phi^{-1}\{F_{y_1}(y_1; \widehat{\boldsymbol{\theta}})\}, \ldots, r^c_{q_n} = \Phi^{-1}\{F_{y_n|(y_1,\ldots,y_{n-1})}(y_n; \widehat{\boldsymbol{\theta}})\},$$

for $i = 1, \ldots, n$. For large n and under the postulated model (1)-(2) $r^c_{q_1}, \ldots, r^c_{q_n}$ follow independent standard normal distributions. So, the normal probability plot may be applied to assess departures from the postulated model as well as to detect possible outlying observations.

8.2 Sensitivity analysis

Assessing the sensitivity of the parameter estimates under perturbations in the model/data is another important tool in regression analysis. Case deletion (see, for instance, Cook and Weisberg 1982) and local influence (Cook 1986) are the most popular procedures. However, dropping observations is not usual in regression models with time series errors, so we will only consider in this section the local influence method, that is the assessing of small perturbations in the model/data on the parameter estimates of model (1)–(2).

We will denote the perturbation vector as $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ restricted to some open subset $\Omega \in \mathcal{R}^n$ and the perturbed penalized log-likelihood function by $\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\omega})$. The no perturbation vector $\boldsymbol{\omega}_0 \in \Omega$ is such that $\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\omega}_0) = \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Using the same influence measure given in Relvas and Paula (2016) the normal curvature in the unitary direction $||\boldsymbol{\ell}|| = 1$ (Cook 1986) is defined as

$$\mathrm{C}_\ell(\boldsymbol{\theta}) = 2|\boldsymbol{\ell}^\top \mathbf{B} \boldsymbol{\ell}|,$$

where $\mathbf{B} = \boldsymbol{\Delta}^\top\{(-\ddot{\mathbf{L}}_p^{\widehat{\theta\theta}})^{-1}\}\boldsymbol{\Delta}$ is a symmetric non-negative definite matrix with $\ddot{\mathbf{L}}_p^{\theta\theta}$ being the observed information matrix (see Appendix B), $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1^\top, \boldsymbol{\Delta}_2^\top, \boldsymbol{\Delta}_3, \boldsymbol{\Delta}_4^\top)^\top$ with $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ being $(r_T \times n)$ and $(r_S \times n)$ matrices with elements $\Delta_{1ji} = \partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\omega})/\partial \boldsymbol{\gamma}_{T_j} \partial \omega_i$ and $\Delta_{2li} = \partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\omega})/\partial \gamma_{S_l} \partial \omega_i$, whereas $\boldsymbol{\Delta}_3$ is $(1 \times n)$ vectors

with elements $\Delta_{3i} = \partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\omega})/\partial\phi\partial\omega_i$ and $\boldsymbol{\Delta}_4$ being $(p \times n)$ matrices with elements $\Delta_{4j'i} = \partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\omega})/\partial\rho_{j'}\partial\omega_i$ evaluated at $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega}_0$, for $i = 1, \ldots, n$, $j = 1, \ldots, r_T, l = 1, \ldots, r_S - 1$ and $j' = 1, \ldots, p$.

Denoting the eigenvalues of $\mathbf{B}$ by $\alpha_1, \ldots, \alpha_n$ with the corresponding normalized eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$, Cook (1986) suggests to study the index plot of $|\mathbf{e}_{\max}|$, relative to $\alpha_{\max}$, to assess the local influence of each observation on the largest direction of the normal curvature. Alternatively, Poon and Poon (1999) proposed the conformal normal curvature that corresponds to a normalized version of the normal curvature but is invariant under scale changes and is defined as

$$\mathrm{B}_\ell(\boldsymbol{\theta}) = \boldsymbol{\ell}^\top \mathbf{B}\boldsymbol{\ell}/\sqrt{\mathrm{tr}(\mathbf{B}^2)},$$

where $0 \leq \mathrm{B}_\ell(\boldsymbol{\theta}) \leq 1$ and $\mathrm{tr}(\mathbf{B}^2) = \sum_{i=1}^n \alpha_i^2$. However, various diagnostic graphs may be derived here. For example, denoting the normalized eigenvalues as $\widehat{\alpha}_{\max} = \widehat{\alpha}_1 \geq \cdots \geq \widehat{\alpha}_k \geq q/\sqrt{n} > \widehat{\alpha}_{k+1} \cdots \widehat{\alpha}_n \geq 0$, where $\widehat{\alpha}_i = \alpha_i/\sqrt{\sum_{j=1}^n \alpha_j^2}$, an aggregate influential measure for all q-influential eigenvectors is defined as

$$m(q)_i = \sqrt{\sum_{j=1}^{k} \widehat{\alpha}_j e_{ji}^2},$$

where e_{ji} denotes the ith element of the jth eigenvector $\mathbf{e}_j$, for $i = 1, \ldots, n$, $j = 1, \ldots, k$ and $q = 0, 1, 2, \ldots$. The index plot of $m(q)_i$ is suggested to reveal those observations that are q-influential, that is, influential for all eigenvectors such that $\mathrm{B}_{e_j} \geq q/\sqrt{n}$.

In particular, if the interest is on evaluating the conformal normal curvature in the direction of the ith observation, represented by the $(n \times 1)$ vector $\mathbf{d}_i$ formed by zeros with 1 at the ith position, one has that

$$\mathrm{B}_{d_i} = \mathrm{B}_i = m^2(0)_i = \sum_{j=1}^{n} \widehat{\alpha}_j e_{ji}^2,$$

that corresponds to the square of the total contribution of the orthonormal eigenvectors. A cutoff criterion suggested by Lee and Xu (2004) considers as most influential those observations such that $\mathrm{B}_i > \bar{\mathrm{B}} + c\mathrm{SD(B)}$, where $\bar{\mathrm{B}}$ and SD(B) denote, respectively, the mean and the standard deviation of $\{\mathrm{B}_i, i = 1, \ldots, n\}$ and c is selected appropriately.

For a particular partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$ with $\boldsymbol{\theta}_1$ defined, for example, as $\boldsymbol{\theta}_1 = \boldsymbol{\gamma}$, $\theta_1 = \phi$ or $\theta_1 = \boldsymbol{\rho}$, one may evaluate the conformal normal curvature

$$\mathrm{B}_\ell(\boldsymbol{\theta}_1) = \boldsymbol{\ell}^\top \mathbf{B}_1 \boldsymbol{\ell}/\sqrt{\mathrm{tr}(\mathbf{B}_1^2)},$$

where $0 \leq \mathrm{B}_\ell(\boldsymbol{\theta}_1) \leq 1$, $\mathbf{B}_1 = \boldsymbol{\Delta}^\top\{(-\ddot{\mathbf{L}}_p^{\widehat{\theta}\widehat{\theta}})^{-1} - \mathbf{G}_p^{\widehat{\theta}_2\widehat{\theta}_2}\}\boldsymbol{\Delta}$ and $\mathbf{G}_p^{\theta_2\theta_2} = \mathrm{blockdiag}\{\mathbf{0}, (\ddot{\mathbf{L}}_p^{\theta_2\theta_2})^{-1}\}$. Again, one may perform the index plot of $\mathrm{B}_i(\boldsymbol{\theta}_1)$ to assess the most

influential observations on all the directions when the subvector $\boldsymbol{\theta}_1$ is considered. In Appendix D the matrices $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$, $\boldsymbol{\Delta}_3$ and $\boldsymbol{\Delta}_4$ are derived for the case-weight perturbation scheme.

9 Simulation studies

In this section we describe a simulation study to assess the large sample behavior of the MPLEs $\hat{f}_T(t)$, $\hat{f}_S(s)$, $\hat{\rho}$ and $\hat{\phi}$ from the iterative process developed in Sect. 4.2. The observations were generated from the additive model

$$y_i = f_T(t_i) + f_S(s_i) + \epsilon_i, \tag{7}$$

where $\epsilon_i = \rho\epsilon_{i-1} + e_i$, $-1 < \rho < 1$ and $e_i \overset{iid}{\sim} S(0, \phi)$, being $f_T(t_i)$ and $f_S(s_i)$ fixed performing the trend and seasonality, respectively, defined as $f_T(t_i) = 2(t_i^3 + \cos(\pi t_i^3))$ with $t_i = i/n$ and $f_S(s_i) = 2\sin(100i/(\pi n))$, for $i = 1, \ldots, n$. The values assigned for the parameters were $\phi = 1$ and $\rho = -0.75, -0.25, 0.25$ and 0.75, with sample sizes $n = 100, 300$ and 500. The data were generated, respectively, under normal, Student-t with $\nu = 3$ degrees of freedom errors and power exponential with shape parameter $k = -0.3$ (shorter-tailed distribution) and $k = 0.5$ (heavier-tailed distribution), named PE_1 and PE_2, respectively, and fitted under the same error distributions. The selected values for the smoothing parameters were $(\lambda_T, \lambda_S) = (700, 2)$, and the number of knots are 12 and 10 to trend and seasonality, respectively. The average estimates of $\hat{\rho}$ and $\hat{\phi}$, the bias and the mean squared error (MSE) were calculated as well as the average values of $\hat{f}_T(t)$ and $\hat{f}_S(s)$. For each scenario 1000 replicates were considered.

From Table 1 one has in the column AIC the proportion of correct classification of the model by the Akaike information criterion discussed in Sect. 7. We notice for all error distributions that the proportion of correct classification increases with the sample size, reaching high values for $n = 500$. Furthermore, we may notice that bias and MSE for $\hat{\rho}$ and $\hat{\phi}$ decrease as the sample size increases. The convergence of $\hat{\rho}$ to the true value is slower under positive correlation. We may notice indication of $\hat{\rho}$ and $\hat{\phi}$ consistency when the generated error distribution agrees with the one assumed in the fitted model. The results under misspecification of the error distribution from the simulated data were omitted in Table 1, but even though one has bias and MSE reduction under some scenarios, it is not clear the consistency of $\hat{\rho}$ and $\hat{\phi}$. Thus, the use of autocorrelation and partial autocorrelation functions as well as diagnostic tools, such as the normal probability plot with the conditional quantile residual, are crucial to assess possible departures from the error distribution assumed in the postulated model.

Figure 1 describes the estimated mean values for the function $f_T(t)$ under normal, Student-t and power exponential error distributions. The convergence to the true curve seems satisfied as the sample size increases, but it appears slower for positive autocorrelation. In Fig. 2 one has the estimated mean values of the function $f_S(s)$ and one may

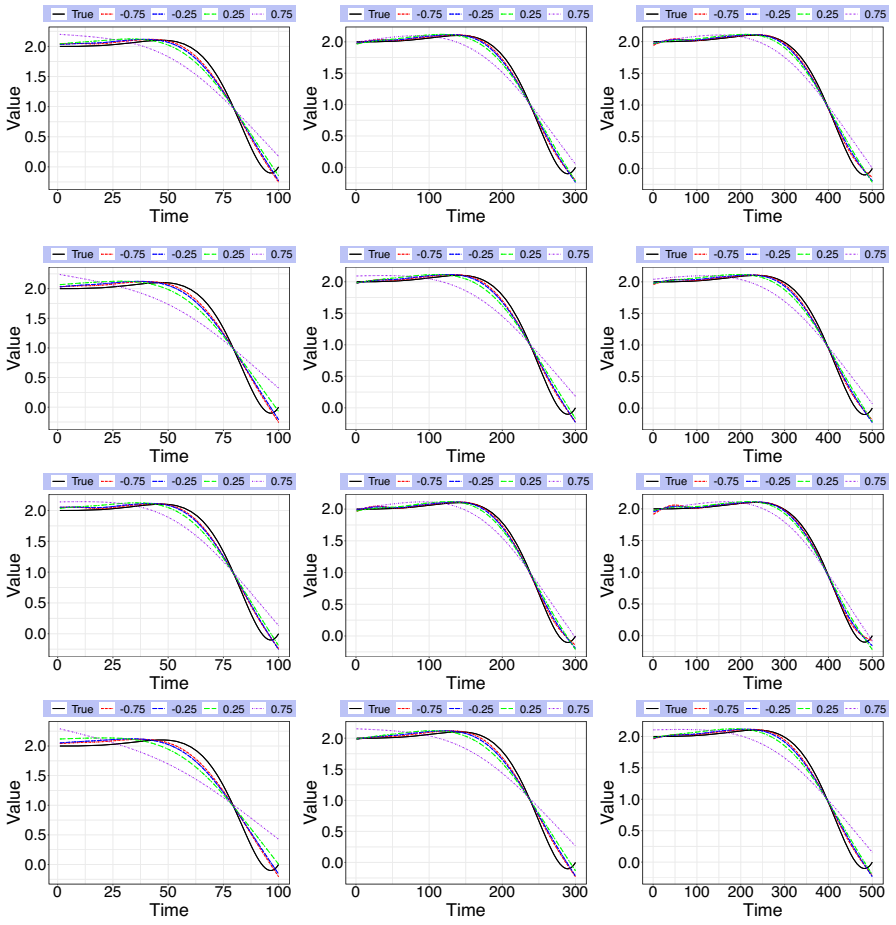

Fig. 1 Graphs of the average estimates for the function $f_T(t)$ under $\phi = 1$ with the autocorrelation coefficients $\rho = -0.75, -0.25, 0.25$ and 0.75 from a simulation study in which the data were generated from model (7) under normal (top), t_3 (second), PE_1(third) and PE_2(bottom) error distributions and fitted under the same error models. Sample sizes $n = 100$, 300 and 500, on the left, middle and right, respectively

notice an excellent behavior for all the scenarios considered, indicating convergence for large n.

10 Daily average temperature of San Francisco

To illustrate the methodology proposed in this paper we will analyze the daily average temperature (in °C) of San Francisco from January 1995 to April 2020, a total of 9252 observations. The 36 missing observations were imputed by the average of the three days before and the three days after the date. The data are from University of Dayton

Table 1 Proportion of correct classification by the Akaike criterion, average estimates $\bar{\hat{\rho}}$, biases and mean squared errors (MSEs) of $\hat{\rho}$, average estimates $\bar{\hat{\phi}}$, biases and mean squared errors (MSEs) of $\hat{\phi}$ from a simulation study in which the data were generated from model (7) under normal, t_3, PE_1 and PE_2 error distributions and fitted under the same error models

Error	ρ	n	AIC	$\hat{\rho}$			$\hat{\phi}$		
				$\bar{\hat{\rho}}$	bias	MSE	$\bar{\hat{\phi}}$	bias	MSE
N	−0.75	100	20.70	−0.7391	0.0109	0.0049	0.9886	−0.0114	0.0186
		300	68.90	−0.7311	0.0189	0.0021	1.0505	0.0505	0.0097
		500	85.70	−0.7314	0.0186	0.0014	1.0626	0.0626	0.0082
	−0.25	100	20.90	−0.2789	−0.0289	0.0104	0.9486	−0.0514	0.0188
		300	69.10	−0.2394	0.0106	0.0036	1.0089	0.0089	0.0066
		500	85.70	−0.2307	0.0193	0.0025	1.0251	0.0251	0.0046
	0.25	100	20.60	0.1609	−0.0891	0.0206	0.9063	−0.0937	0.0259
		300	69.10	0.2267	−0.0233	0.0042	0.9788	−0.0212	0.0066
		500	85.50	0.2433	−0.0067	0.0022	0.9961	−0.0039	0.0037
	0.75	100	26.00	0.6592	−0.0908	0.0183	0.9021	−0.0979	0.0265
		300	69.90	0.7036	−0.0464	0.0046	0.9632	−0.0368	0.0072
		500	86.20	0.7179	−0.0321	0.0023	0.9801	−0.0199	0.0039
t_3	−0.75	100	82.30	−0.7486	0.0014	0.0033	0.9344	−0.0656	0.0455
		300	92.90	−0.7429	0.0071	0.0011	1.0254	0.0254	0.0153
		500	98.20	−0.7426	0.0074	0.0006	1.0386	0.0386	0.0092
	−0.25	100	83.60	−0.3113	−0.0613	0.0095	0.9090	−0.0910	0.0489
		300	94.50	−0.2625	−0.0125	0.0023	0.9884	−0.0116	0.0139
		500	98.10	−0.2507	−0.0007	0.0012	1.0072	0.0072	0.0076
	0.25	100	85.70	0.2361	−0.0139	0.0064	0.8922	−0.1078	0.0504
		300	94.80	0.2414	−0.0086	0.0023	0.9659	−0.0341	0.0153
		500	98.10	0.2472	−0.0028	0.0012	0.9821	−0.0179	0.0078
	0.75	100	86.00	0.7116	−0.0384	0.0072	0.8911	−0.1089	0.0500
		300	95.60	0.7258	−0.0242	0.0020	0.9549	−0.0451	0.0151
		500	98.50	0.7332	−0.0168	0.0010	0.9703	−0.0297	0.0080
PE_1	−0.75	100	89.30	−0.7207	0.0293	0.0061	1.0606	0.0606	0.0194
		300	87.00	−0.7210	0.0290	0.0027	1.1054	0.1054	0.0168
		500	90.90	−0.7210	0.0290	0.0018	1.1037	0.1037	0.0141
	−0.25	100	92.20	−0.3032	−0.0532	0.0090	0.9903	−0.0097	0.0152
		300	91.30	−0.2319	0.0181	0.0033	1.0424	0.0424	0.0071
		500	94.70	−0.2179	0.0321	0.0029	1.0510	0.0510	0.0057
	0.25	100	93.00	0.2372	−0.0128	0.0096	0.9225	−0.0775	0.0211
		300	94.90	0.2569	0.0069	0.0032	0.9922	−0.0078	0.0049
		500	96.70	0.2659	0.0159	0.0021	1.0063	0.0063	0.0029
	0.75	100	90.80	0.6775	−0.0725	0.0143	0.9177	−0.0823	0.0204
		300	95.90	0.7109	−0.0391	0.0036	0.9672	−0.0328	0.0054
		500	97.90	0.7267	−0.0233	0.0016	0.9796	−0.0204	0.0030
PE_2	−0.75	100	17.30	−0.7526	−0.0026	0.0038	0.9352	−0.0648	0.0322

Table 1 continued

Error	ρ	n	AIC	$\hat{\rho}$			$\hat{\phi}$		
				$\bar{\hat{\rho}}$	bias	MSE	$\bar{\hat{\phi}}$	bias	MSE
		300	77.50	−0.7455	0.0045	0.0014	1.0079	0.0079	0.0091
		500	91.50	−0.7426	0.0074	0.0009	1.0188	0.0188	0.0062
	−0.25	100	16.90	−0.3418	−0.0918	0.0138	0.9116	−0.0884	0.0354
		300	78.20	−0.2700	−0.0200	0.0031	0.9850	−0.0150	0.0100
		500	92.00	−0.2537	−0.0037	0.0017	0.9981	−0.0019	0.0055
	0.25	100	15.70	0.2262	−0.0238	0.0087	0.9014	−0.0986	0.0372
		300	79.70	0.2348	−0.0152	0.0030	0.9720	−0.0280	0.0105
		500	93.10	0.2439	−0.0061	0.0018	0.9871	−0.0129	0.0058
	0.75	100	21.30	0.7005	−0.0495	0.0099	0.9065	−0.0935	0.0364
		300	77.70	0.7141	−0.0359	0.0031	0.9716	−0.0284	0.0105
		500	94.20	0.7268	−0.0232	0.0015	0.9822	−0.0178	0.0056

- Environmental Protection Agency Average Daily Temperature Archive.[1] (Kissock 1999)

Similarly to Wood (2017, Sect. 7.7.2) we propose the following additive model for explaining the daily average temperature of San Francisco:

$$\overline{\text{temp}}_i = f_T(\text{time}_i) + f_S(\text{day.of.year}_i) + \epsilon_i, \tag{8}$$

where $f_T(\cdot)$ and $f_S(\cdot)$ are smooth functions approximated by cubic regression and cyclic cubic regression splines, respectively, whereas $\overline{\text{temp}}_i$ denotes the average daily temperature of the ith day, time_i is the ith day, day.of.year_i denotes the day of the year respective to the ith day, and ϵ_i are AR(p) conditional symmetric errors, as defined in Sect. 2, for $i = 1, \ldots, 9252$. Since one has some leap years, the default is day.of.year runs from 1 to 366. We use GCV method to choose the fixed knots number, and for this application we select 80 and 15 knots from time and day of the year, respectively, which were obtained from the quantiles. Homogeneity of seasonality is also assumed over the years.

Under model (8) we hope to control the autocorrelation and seasonality properly and then to detect the trend of the daily average temperature over decades. Figure 3 describes the time series and the empirical distribution, for each month, of the daily average temperature of San Francisco. In order to have a kurtosis flexibility we will fit model (8) to the data under normal error distribution as well as under Student$-t$ error distribution with ν degrees of freedom (heavier-tailed distribution) and under power exponential error distribution with shape parameter k (heavier-tailed distribution for $0 < k < 1$ and shorter-tailed distribution for $-1 < k < 0$), and for different AR(p) structures.

We apply the iterative process described in Sect. 4.2 for a grid of (λ_T, λ_S) values for choosing appropriate smoothing parameter values. Then, we obtain values for AIC

[1] http://academic.udayton.edu/kissock/http/Weather/default.htm.

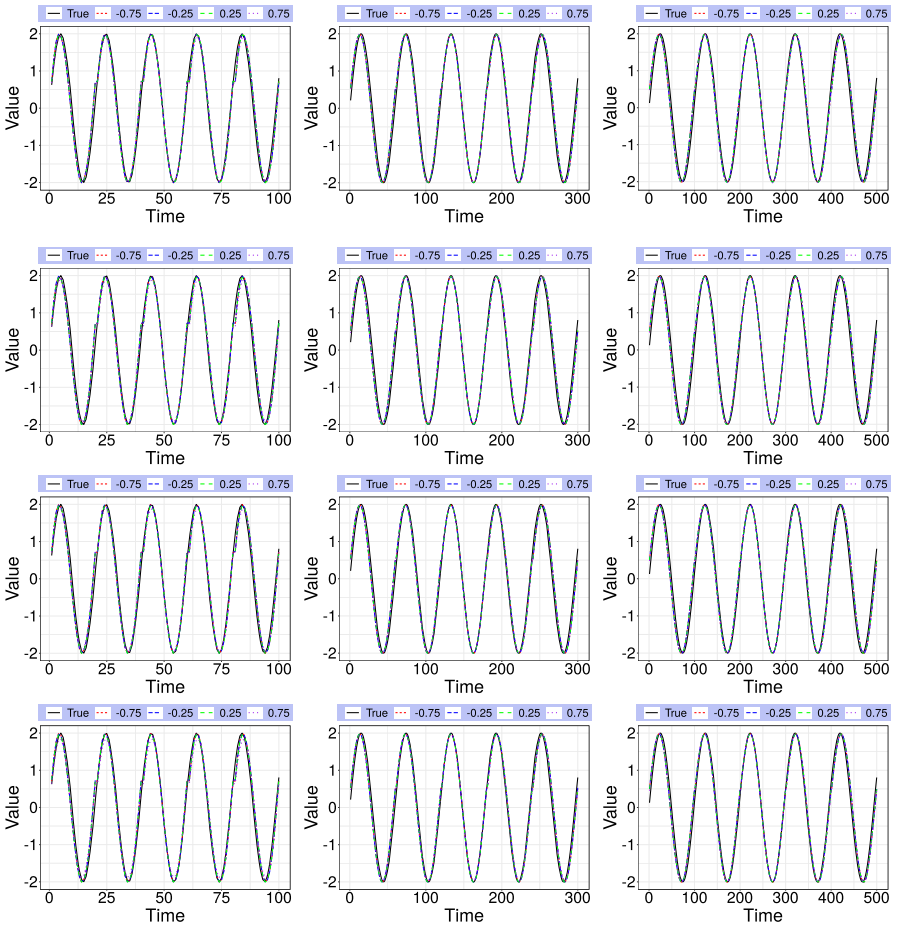

Fig. 2 Graphs of the average estimates for the function $f_S(s)$ under $\phi = 1$ with the autocorrelation coefficients $\rho = -0.75, -0.25, 0.25$ and 0.75 from a simulation study in which the data were generated from model (7) under normal (top), t_3 (second), PE_1(third) and PE_2(bottom) error distributions and fitted under the same error models. Sample sizes $n = 100$, 300 and 500, on the left, middle and right, respectively

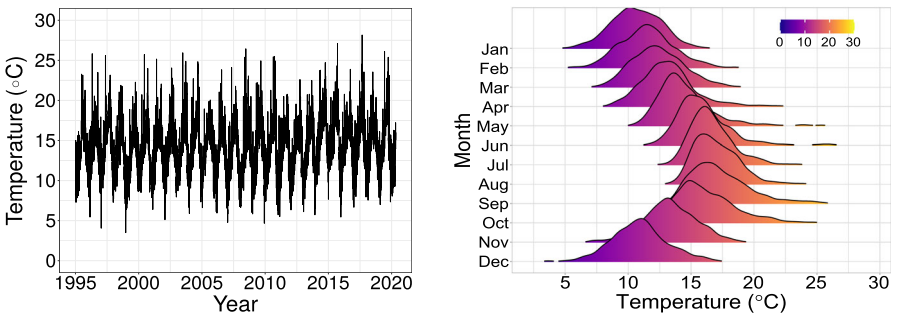

Fig. 3 Time series and the empirical distribution of the daily average temperature (°C) in San Francisco city from January 1995 to April 2020

Table 2 Goodness-of-fit measures from model (8) under normal, Student$-t$ and power exponential errors with AR(1), AR(2) and AR(3) structures fitted to the average daily temperature of San Francisco

AR	Model	AIC($\boldsymbol{\lambda}$)	GCV($\boldsymbol{\lambda}$)	df($\boldsymbol{\lambda}$)	df(λ_T)	df(λ_S)
1	N	32368.2016	1.9360	33.2675	19.9145	11.3530
	t_5	31730.4353	1.1885	36.0641	22.2164	11.8477
	$\text{PE}_{0.6}$	31753.6324	0.5939	42.7275	28.2100	12.5175
2	N	32158.1363	1.8926	36.0248	21.3857	11.6390
	t_5	31560.9024	1.1697	38.5315	23.5117	12.0198
	$\text{PE}_{0.6}$	31579.8760	0.5829	45.0208	29.4649	12.5559
3	N	32130.0399	1.8868	36.3785	20.8378	11.5407
	t_5	31528.2171	1.1654	38.8883	22.9395	11.9488
	$\text{PE}_{0.6}$	31549.3362	0.5809	45.6223	29.0651	12.5572

Table 3 Parametric parameter estimates (approximate standard errors) from model (8) under normal, Student-t and power exponential errors with AR(1), AR(2) and AR(3) structures fitted to the average daily temperature of San Francisco

AR	Model	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\phi}$
1	N	0.7033 (0.0074)	–	–	1.9230 (0.0283)
	t_5	0.7061 (0.0075)	–	–	1.1797 (0.0220)
	$\text{PE}_{0.6}$	0.7141 (0.0068)	–	–	0.5887 (0.0109)
2	N	0.8074 (0.0103)	−0.1495 (0.0103)	–	1.8791 (0.0276)
	t_5	0.7979 (0.0102)	−0.1274 (0.0097)	–	1.1607 (0.0217)
	$\text{PE}_{0.6}$	0.8052 (0.0096)	−0.1246 (0.0094)	–	0.5776 (0.0107)
3	N	0.8161 (0.0104)	−0.1947 (0.0133)	0.0561 (0.0104)	1.8738 (0.0276)
	t_5	0.8056 (0.0103)	−0.1731 (0.0125)	0.0562 (0.0096)	1.1567 (0.0216)
	$\text{PE}_{0.6}$	0.8118 (0.0098)	−0.1684 (0.0119)	0.0540 (0.0098)	0.5757 (0.0107)

and GCV under normal, Student$-t$ and power exponential error distributions with ν and k shape parameters, respectively. These parameters were estimated separately in each fit together with the smoothing parameters. The selected values for the smoothing parameters were $(\lambda_T, \lambda_S) = (500, 5)$. Table 2 describes the AIC and GCV values as well as df(λ_T) and df(λ_S) from model (8) under normal, Student$-t$ (with $\nu = 5$ degrees of freedom) and exponential power (with $k = 0.6$) errors, with AR(1), AR(2) and AR(3) structures, fitted to the daily average temperature of San Francisco. The AR(3) Student-t error model presents the smallest value for AIC, for this reason it was initially selected to explain the daily average temperature in the city of San Francisco. The AR(3) normal error model has the smallest cost (effective degrees of freedom) for estimating $\widehat{\boldsymbol{\gamma}}$, the AR(3) Student$-t$ error model has the intermediate cost and the AR(3) power exponential model has the largest one.

The parametric parameter estimates from the fitted models are described in Table 3. We may notice that the autocorrelation coefficient estimates and their approximate standard errors are very close, but the dispersion parameter estimates are not compa-

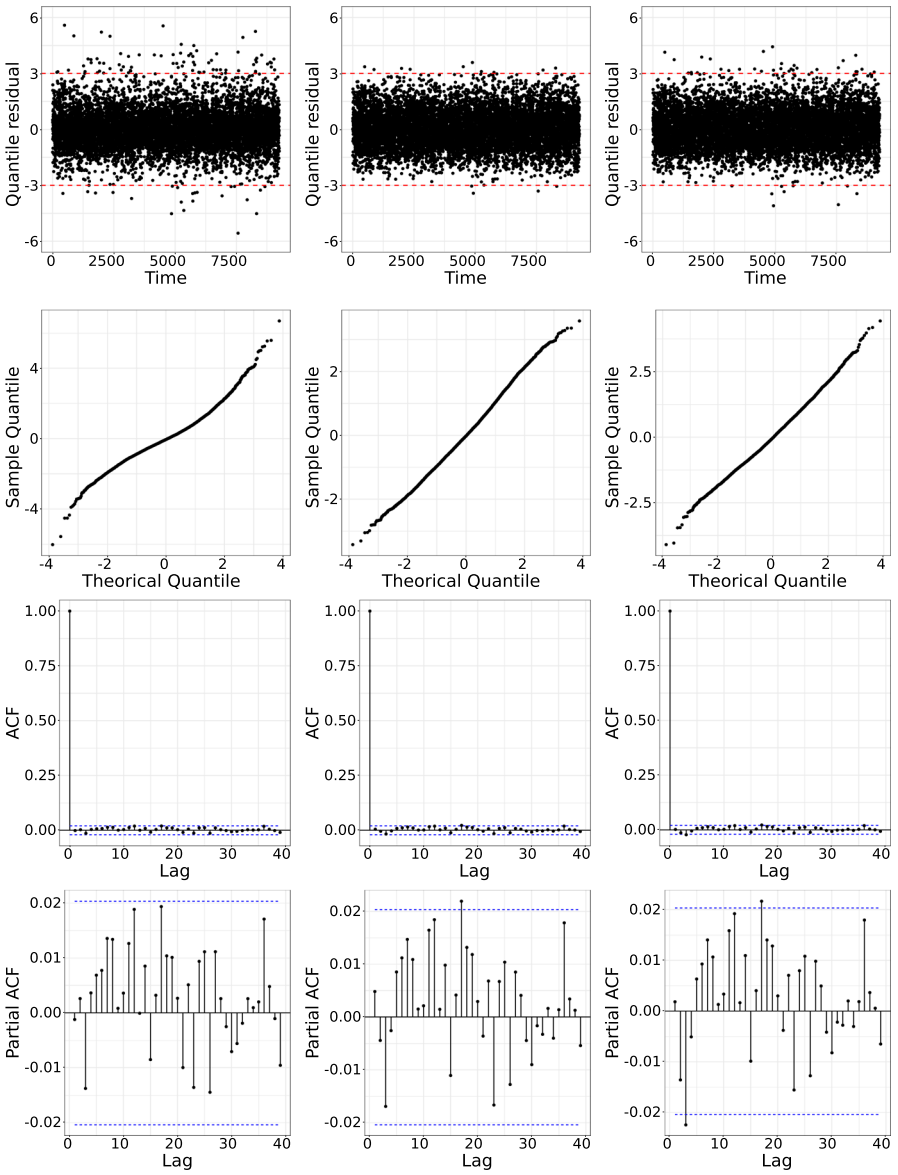

Fig. 4 Residual index plots (top), normal probability plots (second), autocorrelation functions (third) and partial autocorrelation functions (bottom) from model (8) under normal (left), Student-t (middle) and power exponential (right) errors with AR(3) structure fitted to the daily average temperature of San Francisco

rable since they are in different scales. In the sequel we will compare the fitted AR(3) error models according to the residual analysis and sensitivity studies.

Figure 4 describes the index plots and the normal probability plots of the conditional quantile residual from model (8) under normal, Student$-t$ and power exponential

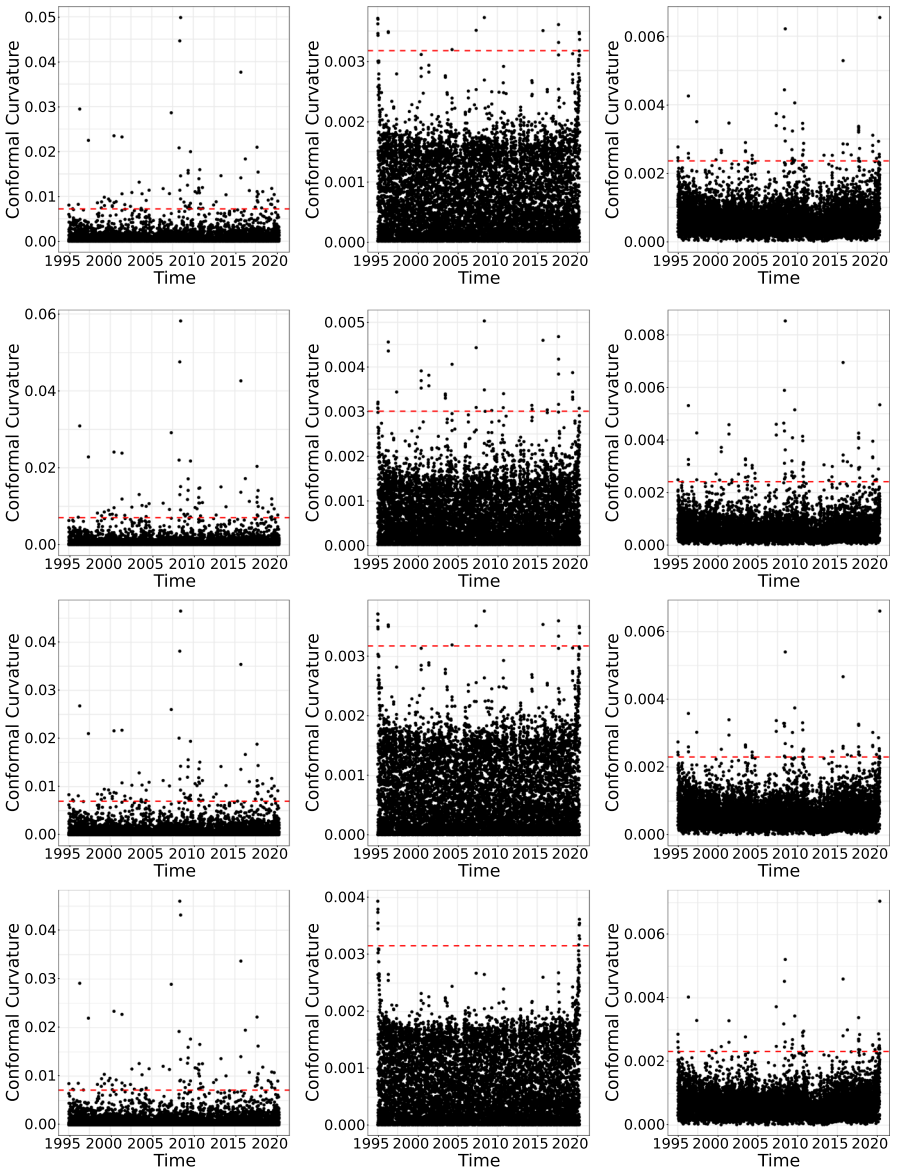

Fig. 5 Index plots of B_i under the case-weight perturbation scheme for $\widehat{\boldsymbol{\theta}}$, $\widehat{\mathbf{f}}$, $\hat{\phi}$ and $\hat{\boldsymbol{\rho}}$ from model (8) under normal (left), Student-t (middle) and power exponential (right) errors with AR(3) structure fitted to the daily average temperature of San Francisco

errors with AR(3) structure fitted to the data. The index plots do not present any tendency neither variability over the time, which indicates that the error variance seems constant and that trend and seasonality were controlled. From the normal probability plots one has indication that the heavier-tailed error models are more suitable to fit the data than the normal one. The autocorrelation functions (ACFs) and the partial

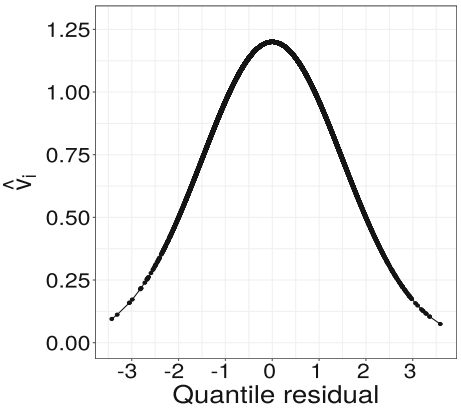

Fig. 6 Graph between the fitted weight $\hat{v}_i$ against the conditional quantile residual from model (8) under Student-t error with $\nu = 5$ degrees of freedom and AR(3) structure fitted to the daily average temperature of San Francisco

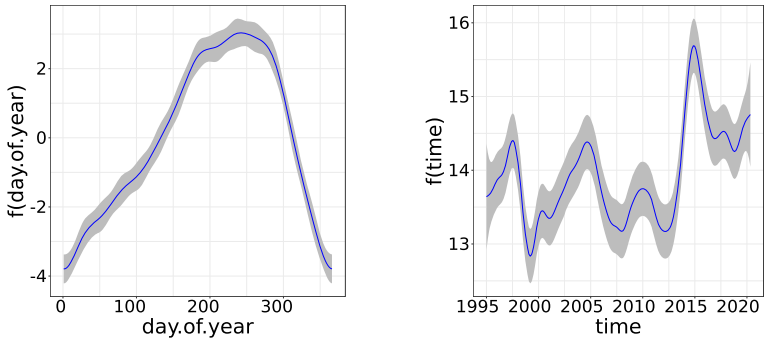

Fig. 7 Pointwise confidence bands for the seasonality components (left) and trend components (right) from model (8) under Student-t error with $\nu = 5$ degrees of freedom and AR(3) structure fitted to the daily average temperature of San Francisco

ACFs from the three series of the conditional quantile residuals confirm the adequacy of AR(3) errors.

Concerning the sensitivity studies Fig. 5 describes the index plots of B_i, under the case-weight perturbation scheme, to assess the conformal local influence on $\widehat{\boldsymbol{\theta}}$, $\widehat{\mathbf{f}}$, $\hat{\phi}$ and $\hat{\boldsymbol{\rho}}$ from model (8) under normal, Student$-t$ and power exponential errors with AR(3) strucuture fitted to the average daily temperature of San Francisco. Similarly to Ibacache-Pulgar et al. (2013) we considered the cutoff value $c = 4$ that generated different cutoff lines for the error models. The graphs indicate that the parameter estimates are less sensitivity under the Student-t error model.

Then, from all the analyzes above the AR(3) Student-t error model with $\nu = 5$ degrees of freedom was selected to explain the daily average temperature of San Francisco. Figure 6, that presents the graph between the fitted weight $\hat{v}_i$ versus the conditional quantile residual from the selected model, confirms the robustness of the parameter estimates from Student-t models against outlying observations (see, for instance, Lucas 1997). Note that, small weights were attributed to large residuals. Finally, Fig. 7 describes the pointwise confidence bands for the seasonality (annual cycle) and for the trend of the daily average temperature of San Francisco from the

chosen model. The left panel indicates a slow growth of the daily average temperature until the peak in September and then a faster decrease with smaller variability. The long term trend appears to be stationary until 2013 with a rise after this year and decreases until April 2020.

11 Concluding remarks

Additive models with conditional AR(p) symmetric errors and linear predictor formed by two nonparametric components approximated by penalized regression splines are proposed in this paper for modeling seasonality and trend in time series. Features of the proposed models are the possibility of robust parameter estimates against outlying observations under heavier-tailed error distributions, kurtosis flexibility and preservation of the time series structure. Various results are derived, such as a backfitting iterative process jointly with a quasi-Newton algorithm for obtaining the MPLEs, inferential procedures, effective degrees of freedom, model selection procedures, residual and sensitivity analyzes and simulation studies. An application for explaining the daily average temperature of San Francisco city is presented for illustrating the procedures developed in the paper. R codes (R Core Team, 2020) developed by the authors for fitting the proposed models to the daily average temperature of San Francisco are available upon request.

Acknowledgements The authors are grateful to the Associate Editor and reviewers for their helpful comments. This study was partially supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001 and CNPq, Brazil.

Appendix A: Penalized score function

Let $\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})$ denote the penalized log-likelihood function for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\gamma}_T^\top, \boldsymbol{\gamma}_S^\top, \phi, \rho_1, \ldots, \rho_p)^\top$. One has that

$$\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda}) = -\frac{n}{2}\log(\phi) + \sum_{i=1}^{n} \log\{g(\delta_i)\} - \frac{\lambda_T}{2}\boldsymbol{\gamma}_T^\top \mathbf{M}_T \boldsymbol{\gamma}_T - \frac{\lambda_S}{2}\boldsymbol{\gamma}_S^\top \mathbf{M}_S \boldsymbol{\gamma}_S,$$

where $\delta_i = \frac{(\epsilon_i - \rho_1\epsilon_{i-1} - \ldots - \rho_p\epsilon_{i-p})^2}{\phi}$, $\epsilon_i = y_i - \mathbf{n}_T(t_i)^\top \boldsymbol{\gamma}_T - \mathbf{n}_S(s_i)^\top \boldsymbol{\gamma}_S$, for $i = 1, \ldots, n$.

The penalized score functions for ϕ and $\boldsymbol{\rho}$ are, respectively, given by

$$\mathrm{U}_p^{\phi} = \frac{\partial \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \phi} = -\frac{1}{2\phi}\left\{n + \sum_{i=1}^{n} 2W_g(\delta_i)\delta_i\right\}$$

and

$$\mathrm{U}_p^{\rho_j} = \frac{\partial \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \rho_j} = -\frac{2}{\phi} \sum_{i=1}^{n} W_g(\delta_i)(\epsilon_i - \rho_1 \epsilon_{i-1} - \ldots - \rho_p \epsilon_{i-p})(\epsilon_{i-j}), \quad (j = 1, \ldots, p).$$

In addition, the derivatives of $\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})$ with respect to γ_{T_j} and γ_{S_l} yield

$$\begin{aligned}
\mathrm{U}_p^{\gamma_{T_j}} &= \frac{\partial \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \gamma_{T_j}} \\
&= -\frac{2}{\phi} \sum_{i=1}^{n} W_g(\delta_i)(\epsilon_i - \rho_1 \epsilon_{i-1} - \cdots - \rho_p \epsilon_{i-p})(n_{T_{ij}} - \rho_1 n_{T_{(i-1)j}} - \cdots - \rho_p n_{T_{(i-p)j}}) \\
&\quad -\lambda_T [\mathbf{M}_T \boldsymbol{\gamma}_T]_j, \quad (j = 1, \ldots, r_T)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{U}_p^{\gamma_{S_l}} &= \frac{\partial \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \gamma_{S_l}} \\
&= -\frac{2}{\phi} \sum_{i=1}^{n} W_g(\delta_i)(\epsilon_i - \rho_1 \epsilon_{i-1} - \cdots - \rho_p \epsilon_{i-p})(n_{S_{il}} - \rho_1 n_{S_{(i-1)l}} - \cdots - \rho_p n_{S_{(i-p)l}}) \\
&\quad -\lambda_S [\mathbf{M}_S \boldsymbol{\gamma}_S]_l, \quad (l = 1, \ldots, r_S - 1),
\end{aligned}$$

where $[\mathbf{M}_T \boldsymbol{\gamma}_T]_j$ and $[\mathbf{M}_S \boldsymbol{\gamma}_S]_l$ denote the jth and lth positions of the vectors $\mathbf{M}_T \boldsymbol{\gamma}_T$ and $\mathbf{M}_S \boldsymbol{\gamma}_S$, respectively.

In matrix notation we obtain

$$\begin{aligned}
\mathbf{U}_p^{\gamma_T} &= \frac{\partial \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\gamma}_T} = \frac{1}{\phi} (\mathbf{A}\mathbf{N}_T)^\top \mathbf{D}_v \mathbf{A} \boldsymbol{\epsilon} - \boldsymbol{\lambda}_T \mathbf{M}_T \boldsymbol{\gamma}_T, \\
\mathbf{U}_p^{\gamma_S} &= \frac{\partial \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\gamma}_S} = \frac{1}{\phi} (\mathbf{A}\mathbf{N}_S)^\top \mathbf{D}_v \mathbf{A} \boldsymbol{\epsilon} - \boldsymbol{\lambda}_S \mathbf{M}_S \boldsymbol{\gamma}_S, \\
\mathrm{U}_p^{\phi} &= \frac{\partial \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \phi} = \frac{1}{2\phi} \mathbf{1}_n^\top (\mathbf{D}_m \mathbf{1}_n - \mathbf{1}_n) \quad \text{and} \\
\mathrm{U}_p^{\rho_j} &= \frac{\partial \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \rho} = -\frac{1}{\phi} (\mathbf{C}_j \boldsymbol{\epsilon})^\top \mathbf{D}_v \mathbf{A} \boldsymbol{\epsilon}, \quad (j = 1, \ldots, p).
\end{aligned}$$

where the quantities $\mathbf{N}_T$, $\mathbf{N}_S$, $\boldsymbol{\epsilon}$, $\mathbf{D}_v$, $\mathbf{D}_m$, $\mathbf{A}$ and $\mathbf{C}_j$ were defined in Sect. 4.

Appendix B: Penalized Hessian matrix

For simplicity of notation we will consider $n_{T_{ij}} = n_{T_j}(t_i)$ and $n_{S_{il}} = n_{S_l}(t_i)$, $(i = 1, \ldots, n)$, $(j = 1, \ldots, r_T)$ and $(l = 1, \ldots, r_S - 1)$.

Consider the parameters $(\gamma_{T_j}, \gamma_{T_h})$ for which we obtain the derivatives

$$\begin{aligned}\ddot{\mathrm{L}}_p^{\gamma_{T_j}\gamma_{T_h}} &= \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \gamma_{T_j} \partial \gamma_{T_h}^\top} \\ &= \frac{4}{\phi} \sum_{i=1}^{n} W_g'(\delta_i)\delta_i (n_{T_{ij}} - \rho_1 n_{T_{(i-1)j}} - \cdots - \rho_p n_{T_{(i-p)j}})(n_{T_{ih}} - \rho_1 n_{T_{(i-1)h}} - \cdots \\ &\quad - \rho_p n_{T_{(i-p)h}}) + \frac{2}{\phi} \sum_{i=1}^{n} W_g(\delta_i)(n_{T_{ij}} - \rho_1 n_{T_{(i-1)j}} - \cdots - \rho_p n_{T_{(i-p)j}}) \times \\ &\quad \times (n_{T_{ih}} - \rho_1 n_{T_{(i-1)h}} - \cdots - \rho_p n_{T_{(i-p)h}}) - \lambda_T \, [\mathbf{M}_T]_{jh}, \quad (j, h = 1, \ldots, r_T).\end{aligned}$$

In matrix notation, we obtain

$$\ddot{\mathrm{L}}_p^{\gamma_T \gamma_T} = \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \gamma_T \partial \gamma_T^\top} = \frac{1}{\phi} \left\{ (\mathbf{A}\mathbf{N}_T)^\top (-\mathbf{D}_v + 4\mathbf{D}_d)(\mathbf{A}\mathbf{N}_T) \right\} - \lambda_T \mathbf{M}_T.$$

Similarly, for the parameter vector $\boldsymbol{\gamma}_S$ one has

$$\ddot{\mathrm{L}}_p^{\gamma_S \gamma_S} = \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \gamma_S \partial \gamma_S^\top} = \frac{1}{\phi} \left\{ (\mathbf{A}\mathbf{N}_S)^\top (-\mathbf{D}_v + 4\mathbf{D}_d)(\mathbf{A}\mathbf{N}_S) \right\} - \lambda_S \mathbf{M}_S.$$

The second derivatives of $\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})$ with respect to ϕ and ρ_j yield

$$\begin{aligned}\ddot{\mathrm{L}}_p^{\phi\phi} &= \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \phi^2} = \frac{n}{2\phi^2} + \frac{2}{\phi^2} \sum_{i=1}^{n} W_g(\delta_i)\delta_i + \frac{1}{\phi^2} \sum_{i=1}^{n} W_g'(\delta_i)\delta_i^2 \\ &= \frac{1}{\phi^2} \left\{ \frac{n}{2} + \boldsymbol{\delta}^\top \mathbf{D}_c \boldsymbol{\delta} - \boldsymbol{\delta}^\top \mathbf{D}_v \mathbf{1}_n \right\}\end{aligned}$$

and

$$\begin{aligned}\ddot{\mathrm{L}}_p^{\rho_j \rho_j} &= \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \rho_j^2} = \frac{4}{\phi} \sum_{i=1}^{n} W_g'(\delta_i)\delta_i \epsilon_{i-j}^2 + \frac{2}{\phi} \sum_{i=1}^{n} W_g(\delta_i)\epsilon_{i-j}^2 \\ &= \frac{1}{\phi} \left\{ (\mathbf{C}_j \boldsymbol{\epsilon})^\top (-\mathbf{D}_v + 4\mathbf{D}_d) (\mathbf{C}_j \boldsymbol{\epsilon}) \right\}, \quad (j = 1, \ldots, p),\end{aligned}$$

where $\mathbf{D}_c = \mathrm{diag}\{c_1, \ldots, c_n\}$ with $c_i = W_g'(\delta_i)$ and $\mathbf{D}_d = \mathrm{diag}\{d_1, \ldots, d_n\}$ with $d_i = W_g'(\delta_i)\delta_i$.

The derivatives of $\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})$ with respect to $(\gamma_{T_j}, \gamma_{S_l})$ yield

$$\ddot{\mathrm{L}}_p^{\gamma_{T_j}\gamma_{S_l}} = \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\gamma_{T_j}\partial\gamma_{S_l}^\top}$$
$$= \frac{1}{\phi}\sum_{i=1}^{n}(n_{T_{ij}} - \rho_1 n_{T_{(i-1)j}} - \cdots - \rho_p n_{T_{(i-p)j}})(n_{S_{il}} - \rho_1 n_{S_{(i-1)l}} - \cdots - \rho_p n_{S_{(i-p)l}}) \times$$
$$\times \left\{4W_g'(\delta_i)\delta_i + 2W_g(\delta_i)\right\}, (j = 1,\ldots,r_T) \quad \text{and} \quad (l = 1,\ldots,r_S - 1).$$

In matrix notation, we obtain

$$\ddot{\mathbf{L}}_p^{\gamma_T\gamma_S} = \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\boldsymbol{\gamma}_T\partial\boldsymbol{\gamma}_S^\top} = \frac{1}{\phi}\left\{(\mathbf{A}\mathbf{N}_T)^\top(4\mathbf{D}_d - \mathbf{D}_v)(\mathbf{A}\mathbf{N}_S)\right\}.$$

For the derivatives of $\mathrm{L}_p(\boldsymbol{\theta},\boldsymbol{\lambda})$ with respect to (γ_{T_j},ϕ) and $(\gamma_{T_j},\rho_{j'})$, we obtain

$$\ddot{\mathrm{L}}_p^{\gamma_{T_j}\phi} = \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\gamma_{T_j}\partial\phi}$$
$$= \phi^{-2}\left\{\sum_{i=1}^{n}(n_{T_{ij}} - \rho_1 n_{T_{(i-1)j}} - \cdots - \rho_p n_{T_{(i-p)j}})(\epsilon_i - \rho_1\epsilon_{i-1} - \cdots - \rho_p\epsilon_{i-p})\right\}$$
$$\times \left\{2W_g(\delta_i) + 2W_g'(\delta_i)\delta_i\right\} \quad \text{and}$$
$$\ddot{\mathrm{L}}_p^{\gamma_{T_j}\rho_{j'}} = \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial\gamma_{T_j}\partial\rho_{j'}}$$
$$= \frac{1}{\phi}\left\{\sum_{i=1}^{n}\left\{4W_g'(\delta_i)\delta_i + 2W_g(\delta_i)\right\}(n_{T_{ij}} - \rho_1 n_{T_{(i-1)j}} - \cdots - \rho_p n_{T_{(i-p)j}})(\epsilon_{i-j'})\right\}$$
$$+\frac{1}{\phi}\left\{2\sum_{i=1}^{n}W_g(\delta_i)(\epsilon_i - \rho_1\epsilon_{i-1} - \cdots - \rho_p\epsilon_{i-p})(n_{T_{(i-j')j}})\right\}, \quad (j' = 1,\ldots,p),$$

which in matrix form may be expressed as

$$\ddot{\mathbf{L}}_p^{\gamma_T\phi} = \frac{1}{\phi}\left\{(\mathbf{A}\mathbf{N}_T)^\top(2\mathbf{D}_d - \mathbf{D}_v)(\mathbf{A}\boldsymbol{\epsilon})\right\}$$

and

$$\ddot{\mathbf{L}}_p^{\gamma_T\rho_{j'}} = \frac{1}{\phi}\left\{(\mathbf{C}_{j'}\boldsymbol{\epsilon})^\top(\mathbf{D}_v - 4\mathbf{D}_d)(\mathbf{A}\mathbf{N}_T) + (\mathbf{C}_{j'}\mathbf{N}_T)^\top\mathbf{D}_v(\mathbf{A}\boldsymbol{\epsilon})\right\}.$$

Similarly, the derivatives of $\mathrm{L}_p(\boldsymbol{\theta},\boldsymbol{\lambda})$ with respect to $(\boldsymbol{\gamma}_S,\phi)$ and $(\boldsymbol{\gamma}_S,\rho_{j'})$ yield

$$\ddot{\mathbf{L}}_p^{\gamma_S\phi} = \frac{1}{\phi}\left\{(\mathbf{A}\mathbf{N}_S)^\top(2\mathbf{D}_d - \mathbf{D}_v)(\mathbf{A}\boldsymbol{\epsilon})\right\}$$

and

$$\ddot{\mathrm{L}}_p^{\gamma_S\rho_{j'}} = \frac{1}{\phi}\left\{(\mathbf{C}_{j'}\boldsymbol{\epsilon})^\top(\mathbf{D}_v - 4\mathbf{D}_d)(\mathbf{A}\mathbf{N}_S) + (\mathbf{C}_{j'}\mathbf{N}_S)^\top\mathbf{D}_v(\mathbf{A}\boldsymbol{\epsilon})\right\}.$$

Finally, for the derivatives of $\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})$ with respect to (ϕ, ρ_j) and $(\rho_j, \rho_{j'})$ we obtain

$$\begin{aligned}
\ddot{\mathrm{L}}_p^{\phi\rho_j} &= \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial\phi\partial\rho_j} \\
&= \frac{1}{\phi^2}\left\{\sum_{i=1}^{n}\epsilon_{i-j}\left\{2W_g'(\delta_i)\delta_i + 2W_g(\delta_i)\right\}(\epsilon_i - \rho_1\epsilon_{i-1} - \cdots - \rho_p\epsilon_{i-p})\right\} \\
&= \frac{1}{\phi^2}\left\{(\mathbf{C}_j\boldsymbol{\epsilon})^\top(\mathbf{D}_v - 2\mathbf{D}_d)(\mathbf{A}\boldsymbol{\epsilon})\right\}, \quad (j = 1, \ldots, p) \text{ and} \\
\ddot{\mathrm{L}}_p^{\rho_j\rho_{j'}} &= \frac{\partial^2 \mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial\rho_j\partial\rho_{j'}} \\
&= -\frac{2}{\phi}\sum_{i=1} n(-\epsilon_{i-j})(\epsilon_{i-j'})\left\{2W_g'(\delta_i)\delta_i + W_g(\delta_i)\right\} \\
&= \frac{1}{\phi}\left\{(\mathbf{C}_{j'}\boldsymbol{\epsilon})^\top(4\mathbf{D}_d - \mathbf{D}_v)(\mathbf{C}_j\boldsymbol{\epsilon})\right\}, \quad (j \neq j' = 1, \ldots, p).
\end{aligned}$$

Appendix C: Penalized Fisher information matrix

Similarly to Relvas and Paula (2016) the penalized Fisher information matrix will be derived from the regularity conditions applied in the regular log-likelihood function $\mathrm{L}(\boldsymbol{\theta})$, namely $\mathrm{E}\{\partial\mathrm{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\} = \mathbf{0}$ and $\mathrm{E}\{\partial^2\mathrm{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top\} = -\mathrm{E}\{[\partial\mathrm{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}]\left[\partial\mathrm{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^\top\right]\}$, and the results $f_g = \mathrm{E}\left\{W_g^2(z^2)z^4\right\}$ and $d_g = \mathrm{E}\left\{W_g^2(z^2)z^2\right\}$ with $z \sim S(0, 1)$.

For the parameter $\boldsymbol{\gamma}_T$ it follows that

$$\mathbf{K}_p^{\gamma_T\gamma_T} = -\mathrm{E}\left\{\frac{1}{\phi}(\mathbf{A}\mathbf{N}_T)^\top(-\mathbf{D}_v + 4\mathbf{D}_d)(\mathbf{A}\mathbf{N}_T) - \lambda_T\mathbf{M}_T\right\}.$$

We may show that $\mathrm{E}(-\mathbf{D}_v + 4\mathbf{D}_d) = -4d_g$ and consequently the penalized Fisher information matrix for $\boldsymbol{\gamma}_T$ is

$$\mathbf{K}_p^{\gamma_T\gamma_T} = \frac{4d_g}{\phi}(\mathbf{A}\mathbf{N}_T)^\top(\mathbf{A}\mathbf{N}_T) + \lambda_T\mathbf{M}_T.$$

Similarly, we obtain

$$\mathbf{K}_p^{\gamma_S\gamma_S} = \frac{4d_g}{\phi}(\mathbf{A}\mathbf{N}_S)^\top(\mathbf{A}\mathbf{N}_S) + \lambda_S\mathbf{M}_S.$$

From the regularity condition $\mathrm{E}\left(\mathrm{U}_{\theta}^{\phi}\right) = 0$ we obtain $\mathrm{E}\{W_g(\delta_i)\delta_i\} = -\frac{1}{2}$, $(i = 1, \ldots, n)$. Then,

$$\mathrm{E}\left\{ \left(\frac{\partial \mathrm{L}_{p_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \phi} \right)^2 \right\} = \mathrm{E}\left\{ \left(\frac{1}{2\phi} + \frac{1}{\phi} W_g(\delta_i)\delta_i \right)^2 \right\}$$
$$= \frac{1}{\phi^2} \left(fg - \frac{1}{4} \right), \ (i = 1, \ldots, n),$$

where $\mathrm{L}_{p_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ denotes the ith element of the penalized log-likelihood function. Therefore, we obtain $\mathrm{K}_p^{\phi\phi} = \frac{n}{4\phi^2}(4 f_g - 1)$.

For the parameter ρ_j one has that

$$\left(\frac{\partial \mathrm{L}_{p_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \rho_j} \right)^2 = \left(\frac{2}{\phi} W_g(\delta_i)(\epsilon_i - \rho_1 \epsilon_{i-1} - \cdots - \rho_p \epsilon_{i-p})(\epsilon_{i-j}) \right)^2 = \frac{4}{\phi} W_g^2(\delta_i)\delta_i \epsilon_{i-j}^2,$$

which implies

$$\mathrm{E}\left(\frac{4}{\phi} W_g^2(\delta_i)\delta_i \epsilon_{i-1}^2 \right) = \mathrm{E}\left[\mathrm{E}\left\{ \frac{4}{\phi}(W_g^2(\delta_i)\delta_i \epsilon_{i-j}^2) \middle| (y_{i-1}, \ldots, y_{i-p}) \right\} \right] = \mathrm{E}\left(\frac{4}{\phi} d_g \epsilon_{i-j}^2 \right).$$

For AR(1) errors, we may express (see, for instance, Judge, 1982) the error ϵ_i as

$$\epsilon_i = e_i + \rho_1 e_{i-1} + \rho_1^2 e_{i-2} + \cdots$$
$$= \sum_{j=0}^{\infty} \rho_1^j e_{i-j},$$

where $j \to \infty$ means that the series has a past over time. Since $|\rho_1| < 1$ the process is stationary. One obtains

$$\mathrm{E}(\epsilon_i) = \sum_{j=1}^{\infty} \rho_1^j \mathrm{E}(e_{i-j}) = 0$$

and

$$\begin{aligned}
\phi_\epsilon \equiv \mathrm{Var}(\epsilon_i) = \mathrm{E}(\epsilon_i^2) &= \mathrm{E}\left\{ (e_i + \rho_1 e_{i-1} + \rho_1^2 e_{i-2} + \ldots)^2 \right\} \\
&= \mathrm{E}\left(e_i^2 + \rho_1^2 e_{i-1}^2 + \rho_1^4 e_{i-2} + \ldots + \rho_1 e_i e_{i-1} + \rho_1^2 e_i e_{i-2} + \ldots \right) \\
&= \mathrm{E}(e_i^2) + \rho_1^2 \mathrm{E}(e_{i-1}^2) + \rho_1^4 \mathrm{E}(e_{i-2}^2) + \ldots \\
&= \phi\xi(1 + \rho_1^2 + \rho_1^4 + \ldots) \\
&= \frac{\phi\xi}{(1 - \rho_1^2)}.
\end{aligned}$$

Then, the Fisher information of ρ_1 reduces to

$$\begin{aligned} \mathrm{K}_p^{\rho_1\rho_1} &= \frac{4}{\phi} d_g \sum_{i=2}^{n} E(\epsilon_{i-1}^2) = \frac{4}{\phi} d_g \sum_{i=1}^{n-1} \mathrm{E}(\epsilon_i^2) = 4 d_g \xi \sum_{i=1}^{n-1} \frac{1}{(1-\rho_1^2)} \\ &= \frac{4 d_g \xi (n-1)}{1-\rho_1^2}. \end{aligned}$$

It is also a simple matter to find autocorrelation in lag s. For example, in lag 1, one has

$$\begin{aligned} \mathrm{Cov}(\epsilon_i, \epsilon_{i-1}) &= E(\epsilon_i \epsilon_{i-1}) \\ &= E\{(\rho_1 \epsilon_{i-1} + e_i)\epsilon_{i-1}\} \\ &= \rho_1 \phi_\epsilon. \end{aligned}$$

Similarly, for lag 2, one obtains

$$\begin{aligned} \mathrm{Cov}(\epsilon_i, \epsilon_{i-2}) &= \mathrm{E}(\epsilon_i \epsilon_{i-2}) \\ &= \mathrm{E}\left[\{\rho_1(\rho_1 \epsilon_{i-2} + e_{i-1}) + e_i\}\epsilon_{i-2}\right] \\ &= \rho_1^2 \phi_\epsilon, \end{aligned}$$

and the covariance between l periods of two errors is given by

$$\mathrm{Cov}(\epsilon_i, \epsilon_{i-l}) = \mathrm{E}(\epsilon_i \epsilon_{i-l}) = \mathrm{E}(\epsilon_{i+l}\epsilon_i) = \frac{\rho_1^l \phi \xi}{1-\rho_1^2}.$$

Thus, matrix $\boldsymbol{\Upsilon}_1$ may be constructed.

For AR(2) errors ϵ_i may be expressed as

$$\begin{aligned} \epsilon_i &= \rho_1 \epsilon_{i-1} + \rho_2 \epsilon_{i-2} + e_i \\ &= \rho_1(\rho_1 \epsilon_{i-2} + \rho_2 \epsilon_{i-3} + e_{i-1}) + \rho_2(\rho_1 \epsilon_{i-3} + \rho_2 \epsilon_{i-4} + e_{i-2}) + e_i \\ &= e_i + \rho_1 e_{i-1} + \rho_2 e_{i-2} + \ldots, \end{aligned} \tag{9}$$

where $\mathrm{E}(e_i) = 0$, $\mathrm{E}(e_i e_{i'}) = 0$, for $i \neq i'$, and $\mathrm{E}(e_i^2) = \phi\xi$, thereby $\mathrm{E}(\epsilon_i) = 0$. This process is stationary if $\rho_1 + \rho_2 < 1$, $\rho_2 - \rho_1 < 1$ and $-1 < \rho_2 < 1$. According to Judge et al. (1985) and Fox (2015), the elements of the covariance matrix $\boldsymbol{\Upsilon}_2$ may be found from the variance

$$\phi_\epsilon \equiv \mathrm{Var}(\epsilon_i) = \mathrm{E}(\epsilon_i^2) = \phi\xi \frac{(1-\rho_2)}{(1+\rho_2)\{(1-\rho_2)^2 - \rho_1^2\}}.$$

Multiplying (9) by ϵ_{i-1} and taking expectation, one obtains

$$\begin{aligned}\mathrm{Cov}(\epsilon_i, \epsilon_{i-1}) &= \rho_1 \mathrm{E}(\epsilon_{i-1}^2) + \rho_2 \mathrm{E}(\epsilon_{i-1}\epsilon_{i-2}) \\ &= \rho_1 \phi_\epsilon + \rho_2 \mathrm{Cov}(\epsilon_i, \epsilon_{i-1}),\end{aligned}$$

and since $\mathrm{E}(\epsilon_{i-1}^2) = \phi_\epsilon$ and $\mathrm{E}(\epsilon_{i-1}\epsilon_{i-2}) = \mathrm{Cov}(\epsilon_{i-1}, \epsilon_{i-2}) = \mathrm{Cov}(\epsilon_i, \epsilon_{i-1})$, solving for autocovariance one obtains

$$\phi_1 \equiv \mathrm{Cov}(\epsilon_i, \epsilon_{i-1}) = \frac{\rho_1}{1-\rho_2}\phi_\epsilon.$$

Similarly, for $l > 1$,

$$\begin{aligned}\phi_l \equiv \mathrm{Cov}(\epsilon_i, \epsilon_{i-l}) &= \rho_1 \mathrm{E}(\epsilon_{i-1}\epsilon_{i-l}) + \rho_2 \mathrm{E}(\epsilon_{i-2}\epsilon_{i-l}) \\ &= \rho_1 \phi_{l-1} + \rho_2 \phi_{l-2},\end{aligned}$$

so we may find the the autocovariance recursively. For example, for $l = 2$, one has

$$\begin{aligned}\phi_2 &= \rho_1 \phi_1 + \rho_2 \phi_0 \\ &= \rho_1 \phi_1 + \rho_2 \phi_\epsilon,\end{aligned}$$

where $\phi_0 = \phi_\epsilon$, and for $l = 3$,

$$\phi_3 = \rho_1 \phi_2 + \rho_2 \phi_1.$$

In general for AR(p) errors one may write

$$\mathrm{K}_p^{\rho_j \rho_j} = \frac{4d_g}{\phi} \sum_{i=j+1}^{n} \mathrm{E}\left(\epsilon_{i-j}^2\right) = \frac{4d_g}{\phi} \sum_{i=1}^{n-j} \mathrm{E}\left(\epsilon_i^2\right) = \frac{4d_g(n-j)}{\phi}\phi_\epsilon,$$

and the Fisher information for $(\rho_j, \rho_{j'})$, $j \neq j' = 1, \ldots, p$, is given by

$$\begin{aligned}\mathrm{K}_p^{\rho_j \rho_j} &= \mathrm{E}\left(-\frac{\partial^2 \mathrm{L}_{p_i}(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \rho_j \partial \rho_{j'}}\right) \\ &= \mathrm{E}\left\{\frac{4}{\phi} W_g'(\delta_i)\delta_i(\epsilon_{i-j})(\epsilon_{i-j'}) + \frac{2}{\phi} W_g(\delta_i)(\epsilon_{i-j})(\epsilon_{i-j'})\right\} \\ &= \mathrm{E}\left\{\frac{1}{\phi}\left(4W_g'(\delta_i)\delta_i + 2W_g(\delta_i)\right)(\epsilon_{i-j})(\epsilon_{i-j'})\right\} \\ &= -\frac{1}{\phi}\mathrm{E}\left[\mathrm{E}\left\{\left(4W_g'(\delta_i)\delta_i + 2W_g(\delta_i)\right)(\epsilon_{i-j})(\epsilon_{i-j'}) \mid y_{i-1}, \ldots, y_{i-p}\right\}\right] \\ &= \frac{4d_g}{\phi}\mathrm{E}(\epsilon_{i-j}\epsilon_{i-j'}).\end{aligned}$$

Considering $j' < j$, we may write the Fisher information for $(\rho_j, \rho_{j'})$, $j \neq j'$, as follows

$$\mathrm{K}_p^{\rho_j \rho_{j'}} = \frac{4d_g}{\phi} \sum_{i=j+1}^{n+j'} \mathrm{E}(\epsilon_{i-j}\epsilon_{i-j'}) = \frac{4d_g}{\phi} \sum_{i=1}^{n-j+j'} \mathrm{E}(\epsilon_i \epsilon_{i+j-j'}) = \frac{4d_g(n-j+j')}{\phi}\phi_{j-j'}.$$

The expression for the Υ_p becomes progressively more complicated. A general expression is given in Wise (1955).

The penalized Fisher information matrix for $(\boldsymbol{\gamma}_T^\top, \boldsymbol{\gamma}_S^\top)^\top$ is given by

$$\mathbf{K}_p^{\gamma_T \gamma_S} = -\mathrm{E}\left\{\frac{1}{\phi}(\mathbf{A}\mathbf{N}_T)^\top(4\mathbf{D}_d - \mathbf{D}_v)(\mathbf{A}\mathbf{N}_S)\right\} = \frac{4d_g}{\phi}\left\{(\mathbf{A}\mathbf{N}_T)^\top(\mathbf{A}\mathbf{N}_S)\right\},$$

and we can see that $\boldsymbol{\gamma}_T$ and $\boldsymbol{\gamma}_S$ are not orthogonal.

From the properties of the symmetric distributions we have that

$$\mathrm{E}\left\{\left(W_g(\delta_i) + W_g'(\delta_i)\delta_i\right)(\epsilon_i - \rho_1\epsilon_{i-1} - \cdots \rho_p\epsilon_{i-p}) \mid y_{i-1}, \ldots, y_{i-p}\right\} = 0.$$

Then, we may obtain the following penalized Fisher information matrices:

$$\begin{aligned}
\mathbf{K}_p^{\phi\gamma_T} &= -\mathrm{E}\left[\frac{1}{\phi^2}\left\{(\mathbf{A}\mathbf{N}_T)^\top(-\mathbf{D}_v + 2\mathbf{D}_d)(\mathbf{A}\boldsymbol{\epsilon})\right\}\right] = \mathbf{0}, \\
\mathbf{K}_p^{\phi\gamma_S} &= -\mathrm{E}\left[\frac{1}{\phi^2}\left\{(\mathbf{A}\mathbf{N}_S)^\top(-\mathbf{D}_v + 2\mathbf{D}_d)(\mathbf{A}\boldsymbol{\epsilon})\right\}\right] = \mathbf{0} \text{ and} \\
\mathrm{K}_p^{\phi\rho_j} &= -\mathrm{E}\left[\frac{1}{\phi^2}\left\{(\mathbf{C}_j\boldsymbol{\epsilon})^\top(\mathbf{D}_v - 2\mathbf{D}_d)(\mathbf{A}\boldsymbol{\epsilon})\right\}\right] = 0, \quad \forall j = 1, \ldots, p.
\end{aligned}$$

From $\mathrm{E}(\mathbf{U}_p^{\gamma_T}) = \mathbf{0}$ it follows that

$$\mathrm{E}\left\{\frac{1}{\phi}(\mathbf{A}\mathbf{N}_T)^\top\mathbf{D}_v\mathbf{A}\boldsymbol{\epsilon}\right\} = \mathbf{0},$$

then $\mathrm{E}(\mathbf{D}_v\mathbf{A}\boldsymbol{\epsilon}) = \mathbf{0}$, so we may obtain

$$\begin{aligned}
\mathbf{K}_p^{\rho_j\gamma_T} &= -\frac{1}{\phi}\mathrm{E}\left\{(\mathbf{C}_j\boldsymbol{\epsilon})^\top(\mathbf{D}_v - 4\mathbf{D}_d)(\mathbf{A}\mathbf{N}_T) + (\mathbf{C}_j\mathbf{N}_T)^\top\mathbf{D}_v(\mathbf{A}\boldsymbol{\epsilon})\right\} \\
&= -\frac{1}{\phi}\mathrm{E}\left\{(\mathbf{C}_j\boldsymbol{\epsilon})^\top(\mathbf{D}_v - 4\mathbf{D}_d)(\mathbf{A}\mathbf{N}_T)\right\},
\end{aligned}$$

and since $\mathrm{E}\{(\mathbf{C}_j\boldsymbol{\epsilon})^\top\} = \mathbf{0}$ we have that

$$\mathbf{K}_p^{\rho_j\gamma_T} = -\mathrm{E}\left[\mathrm{E}\left\{\frac{1}{\phi}\left((\mathbf{C}_j\boldsymbol{\epsilon})^\top(\mathbf{D}_v - 4\mathbf{D}_d)\mathbf{A}\mathbf{N}_T\right)\right\} \mid y_{i-1} - \cdots - y_{i-p}\right]$$

$$= \mathrm{E}\left\{\frac{4d_g}{\phi}(\mathbf{C}_j\boldsymbol{\epsilon})^\top(\mathbf{A}\mathbf{N}_T)\right\} = \mathbf{0}.$$

Similarly, we may show that $\mathbf{K}_p^{\rho_j\gamma_S} = \mathbf{0}$.

Appendix D: Case-weight perturbation scheme

Consider the attributed weights in the penalized log-likelihood function as

$$\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\omega}) = \mathrm{L}_p(\boldsymbol{\theta} \mid \boldsymbol{\omega}) - \frac{\lambda_T}{2}\boldsymbol{\gamma}_T^\top\mathbf{M}_T\boldsymbol{\gamma}_T - \frac{\lambda_S}{2}\boldsymbol{\gamma}_S^\top\mathbf{M}_S\boldsymbol{\gamma}_S,$$

where $\mathrm{L}_p(\boldsymbol{\theta} \mid \boldsymbol{\omega}) = \sum_{i=1}^n \omega_i \mathrm{L}_i(\boldsymbol{\theta})$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ is the vector of weights, with $0 \leq \omega_i \leq 1$. In this case the vector of no perturbation is given by $\boldsymbol{\omega}_0 = \mathbf{1}_n$.

For this perturbation scheme we obtain

$$\begin{aligned}
\boldsymbol{\Delta}_1 &= \frac{1}{\hat{\phi}}(\widehat{\mathbf{A}}\mathbf{N}_T)^\top\widehat{\mathbf{D}}_v\widehat{\mathbf{D}}_{(\mathrm{A}_\epsilon)},\\
\boldsymbol{\Delta}_2 &= \frac{1}{\hat{\phi}}(\widehat{\mathbf{A}}\mathbf{N}_S)^\top\widehat{\mathbf{D}}_v\widehat{\mathbf{D}}_{(\mathrm{A}_\epsilon)},\\
\boldsymbol{\Delta}_3 &= \frac{1}{2\hat{\phi}}\mathbf{1}_n^\top\left(\widehat{\mathbf{D}}_m - \mathbf{I}_n\right) \text{ and}\\
\boldsymbol{\Delta}_{4_j} &= \frac{1}{\hat{\phi}}(\widehat{\mathbf{C}}_j\widehat{\boldsymbol{\epsilon}})^\top\widehat{\mathbf{D}}_v\widehat{\mathbf{D}}_{(\mathrm{A}_\epsilon)}, \quad \forall j = 1, \ldots, p,
\end{aligned}$$

where $\mathbf{D}_v = \mathrm{diag}\{v_1, \ldots, v_n\}$ with $v_i = -2W_g(\delta_i)$, $\mathbf{D}_m = \mathrm{diag}\{m_1, \ldots, m_n\}$ with $m_i = v_i\,\delta_i$, for $i = 1, \ldots, n$, $\epsilon_0 = 0$ and $\mathbf{D}_{(\mathrm{A}_\epsilon)}$ is a diagonal matrix with elements given by $\mathbf{A}\boldsymbol{\epsilon}$ evaluated at $\widehat{\boldsymbol{\theta}}$. In addition,

$$\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{N}_T\boldsymbol{\gamma}_T - \mathbf{N}_S\boldsymbol{\gamma}_S \text{ and } \delta_i = \frac{(\epsilon_i - \rho_1\epsilon_{i-1} - \ldots - \rho_p\epsilon_{i-p})^2}{\phi}.$$

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) International symposium on information theory. Akademiai Kiado Budapest, Hungary, pp 267–281

Barros M, Paula GA (2019) Discussion of Birnbaum-Saunders distributions: a review of models, analysis and applications. Appl Stoch Models Bus Ind 35:96–99

Byrd RH, Lu P, Nocedal J, Zhu C (1995) A limited memory algorithm for bound constrained optimization. SIAM J Sci Comput 16:1190–1208

Cao CZ, Lin JG, Zhu LX (2010) Heteroscedasticity and/or autocorrelation diagnostics in nonlinear models with AR(1) and symmetrical errors. Stat Pap 51:813–836

Cook RD (1986) Assessment of local influence. J R Stat Soc B 48:133–169

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman and Hall, London

Cleveland WS, McRae JE, Terpenning I (1990) STL: a seasonal-trend decomposition. J Off Stat 6:3–73

Cysneiros FJA, Paula GA (2005) Restricted methods in symmetrical linear regression models. Comput Stat Data Anal 49:689–708
Davidon WC (1991) Variable metric method for minimization. SIAM J Optim 1:1–17
Dunn PK, Smyth GK (1996) Randomized quantile residuals. J Comput and Graph Stat 5:236–244
Efron B, Hinkley DV (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika 65:457–487
Eilers PH, Marx BD (1996) Flexible smoothing with B-splines and penalties. Stat Sci 11:89–102
Fang KT, Kotz S, Ng KW (1990) Symmetric multivariate and related distributions. Chapman and Hall, London
Fox J (2015) Applied regression analysis and generalized linear models, 3rd edn. Sage Publications, London
Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman and Hall/CRC, London
Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman and Hall/CRC, London
Huang L, Jiang H, Wang H (2019) A novel partial-linear single-index model for time series data. Comput Stat Data Anal 134:110–122
Huang L, Xia Y, Qin X (2016) Estimation of semivarying coefficient time series models with ARMA errors. Ann Stat 44:1618–1660
Ibacache-Pulgar G, Paula GA, Cysneiros FJA (2013) Semiparametric additive models under symmetric distributions. TEST 22:103–121
Judge GG, Griffiths WE, Hill RC, Lutkepohl H, Lee TC (1985) The theory and practice of econometrics, 2nd edn. Wiley, New York
Kissock JK (1999) UD EPA Average Daily Temperature Archive, http://academic.udayton.edu/kissock/http/Weather/default.htm. Accessed 20 Feb 2021
Lancaster P, Salkauskas K (1986) An introduction curve and surface fitting. Academic Press, London
Lee SY, Xu L (2004) Influence analyses of nonlinear mixed-effects models. Comput Stat Data Anal 45:321–341
Liu JM, Chen R, Yao Q (2010) Nonparametric transfer function models. J Econom 157:151–164
Liu S (2004) On diagnostics in conditionally heteroscedastic time series models under elliptical distributions. J Appl Probab 41A:393–405
Lucas A (1997) Robustness of the student-t based M-estimator. Commun Stat Theory Methods 26:1165–1182
Mittelhammer RC, Judge GG, Miller DJ (2000) Econometric foundations. Cambridge University Press, New York
Paula GA, Medeiros MJ, Vilca-Labra FE (2009) Influence diagnostics for linear models with first-order autoregressive elliptical errors. Stat Probab Lett 79:339–346
Poon WY, Poon YS (1999) Conformal normal curvature and assessment of local influence. J R Stat Soc B 61:51–61
R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.Rproject.org. Accessed 10 Jan 2021
Relvas CEM, Paula GA (2016) Partially linear models with first-order autoregressive symmetric errors. Stat Pap 57:795–825
Schwarz GE (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Vanegas LH, Paula GA (2016) An extension of log-symmetric regression models: R codes and applications. J Stat Comp Simul 86:1709–1735
Wise J (1955) The autocorrelation function and the spectral density function. Biometrika 42:151–159
Wood SN (2017) Generalized additive models: an introduction with R, 2nd edn. Chapman and Hall/CRC, London