Statistical Papers
https://doi.org/10.1007/s00362-022-01300-4

REGULAR ARTICLE

# Generalized log-gamma additive partial linear models with P-spline smoothing

**Carlos A. Cardozo**[1] · **Gilberto A. Paula**[2] · **Luiz H. Vanegas**[3]

Received: 5 February 2021 / Revised: 22 November 2021 / Accepted: 14 February 2022


## Abstract

In this paper additive partial linear models with generalized log-gamma errors and P-spline smoothing are proposed for uncensored data. This class derived from the generalized gamma distribution contains various continuous asymmetric distributions to the right and to the left with domain on the real line and has the normal distribution as a particular case. The location parameter is modeled in a semiparametric way so that one has a generalized gamma accelerated failure time additive partial linear model. A joint iterative process is derived, that combines the penalized Fisher scoring algorithm for estimating the parametric and nonparametric regression coefficients and a quasi-Newton procedure for obtaining the scale and shape estimates. Discussions on the inferential aspects of the former estimators as well as on the derivation of the effective degrees of freedom are given. Diagnostic procedures are also proposed, such as residual analysis and sensitivity studies based on the local influence approach. Simulation studies are performed to assess the empirical distributions of the parametric and nonparametric estimators and a real data set on personal injury insurance claims made in Australia from January 1998 to June 1999 is analyzed by the methodology developed through the paper. Technical results, tables, graphs, R codes and the data set used in the application are presented as Supplementary Materials.

**Keywords** AFT models · Asymmetric data · B-splines · Diagnostic procedures · Insurance data · P-GAM · Semiparametric models

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

Carlos A. Cardozo
carlos.cardozo@docentes.umb.edu.co

Luiz H. Vanegas
lhvanegasp@unal.edu.co

[1] Vicerrectoría de Investigaciones, Universidad Manuela Beltran, Bogotá, Colombia

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

[3] Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia

Published online: 12 March 2022

**Mathematics Subject Classification** 60E05 · 62N99 · 62G10

## 1 Introduction

The generalized log-gamma family has its most remote ancestor in the works by the Italian economist (Amoroso 1925), on the distribution of the income. Almost 40 years after, the generalized gamma distribution was derived with the work by Stacy (1962). It was an effort to unifying the most important distributions in reliability and survival analysis, such as exponential, Weibull, Inverse Weibull, gamma, log-normal and Rayleigh among others. Some estimation properties and a new parametrization of the generalized log-gamma distribution were studied in the 60s and 70s of the past century by Stacy and Mihram (1965), Hager and Bain (1970), Prentice (1974) and Lawless (1980, 2003).

Regression models with generalized log-gamma distributed errors have also been studied in the last decades. As a special case of such models is the extreme value (Type I) regression models, which was studied, for instance, by Paula and Rojas (1997) in the context of one-sided tests under uncensored observations. Ortega et al. (2003) presented some estimation procedures and derived some diagnostic quantities based on the local influence approach under the presence of censored observations and Fabio et al. (2012) assumed a generalized log-gamma distribution for the random effect in random intercept Poisson models, whereas Ortega et al. (2009) proposed the log-gamma regression model with cure fraction giving emphasis to the local influence approach. More recently, Agostinelli et al. (2014) derived robust and consistent estimators for the three parameters of the distribution and developed the library `robustlogamma` (Agostinelli et al. 2017) in the software `R` (R Core Team 2019) for fitting log-gamma distributions.

Although this family is very flexible to allow distributions to the left and to the right and has great potential of application in reliability studies as well as in survival analysis, there are few works concerning this family in the context of semiparametric analysis, particularly in GAMLSS (see, for instance, Stasinopoulos et al. 2017). So, the aim of this paper is to propose a general framework for additive partial linear models with generalized log-gamma errors and P-spline smoothing for uncensored data, that corresponds to a generalized gamma accelerated failure time (AFT) model in which the location parameter is modeled in a semiparametric way.

The paper is organized as follows. In Sect. 2 a review on the generalized gamma distribution with various properties and graphs on its probability density function (p.d.f.), hazard function and quantile function is given. The main results related with the generalized log-gamma additive partial linear models are given in Sect. 3, such as a review on the generalized log-gamma distributions and P-spline smoothing. In Sect. 4, the penalized score function and the penalized Fisher information matrix are derived for the parametric and nonparametric components. A joint iterative process that combines the penalized Fisher scoring algorithm for estimating the parametric and nonparametric regression coefficients and a quasi-Newton procedure for obtaining the scale and shape estimates is developed. Some discussions on the asymptotic properties of the former estimators are also given. An appropriate smoother and the effective degrees of freedom are derived in Sect. 5. Diagnostic procedures are proposed in Sect. 6, such as residual analysis based on the quantile residual and sensitivity studies based on local influence. Simulation studies to assess the large sample behavior of the parametric and nonparametric estimators are presented in Sect. 7. In Sect. 8 a real data set on personal injury insurance claims made in Australia, from January 1998 to June 1999 and described in de Jong and Heller (2008), is analyzed by the procedures developed through the paper. The last section deals with some concluding remarks and various technical results, tables, graphs, R codes and the data set used in the application are presented as supplementary materials.

## 2 Generalized gamma distribution

### 2.1 Definition and properties

Let $T$ be a random variable following the generalized gamma distribution. The p.d.f. of $T$ (see, for instance, Lawless 1980) is given by

$$f_T(t;\eta,\sigma,\lambda) = \begin{cases} \frac{C(\lambda)}{t\sigma}\left(\frac{t}{\eta}\right)^{\frac{1}{\sigma\lambda}} e^{-\frac{1}{\lambda^2}\left(\frac{t}{\eta}\right)^{\frac{\lambda}{\sigma}}} & \text{if } \lambda \neq 0 \\ \frac{1}{\sigma t\sqrt{2\pi}} e^{-\frac{1}{\sigma^2}\{\log(t)-\eta\}^2} & \text{if } \lambda = 0, \end{cases},$$

where $t > 0$ and $C(\lambda) = \frac{|\lambda|}{\Gamma(\lambda^{-2})}(\lambda^{-2})^{\lambda^{-2}}$. The location, scale and shape parameters are $\eta > 0$, $\sigma > 0$ and $\lambda \in \mathbb{R}$, respectively, and we will denote $T \sim \text{GG}(\eta,\sigma,\lambda)$. This family of distributions is very flexible as its special cases include log-normal ($\lambda = 0$), exponential ($\eta = \sigma = \lambda = 1$), chi-squared ($\eta = 1, \sigma = \lambda = \sqrt{2}$), gamma ($\sigma = \lambda$), inverse gamma ($\lambda = -\sigma$), Rayleigh ($\lambda = 1, \sigma = 1/2$), Weibull ($\lambda = 1$), inverse Weibull ($\lambda = -1$), Ammag ($\lambda = 1/\sigma$), inverse Ammag ($\lambda = -1/\sigma$), and half-normal ($\eta = 1, \lambda = 1/\sigma = \sqrt{2}$). In addition, if the distribution of $Z$ belongs to the family of exponential power distributions with location parameter $\tau$ and scale parameter $\omega$ (that is, the density function of $Z$ is proportional to $\exp\{-\frac{1}{2}(|z-\tau|/\omega)^{2\nu}\}$ for some $\nu > 0$), which includes the normal ($\nu = 1$), Laplace or double exponential ($\nu = \frac{1}{2}$) and uniform ($\nu \to \infty$) as special cases, then $|Z-\tau|/\omega \sim \text{GG}(\nu^{-\frac{1}{2\nu}}, 1/\sqrt{2\nu}, \sqrt{2\nu})$ and $(Z-\tau)^2/\omega^2 \sim \text{GG}(\nu^{-\frac{1}{\nu}}, \sqrt{2/\nu}, \sqrt{2\nu})$. Figure 1 describes the p.d.f. of the $\text{GG}(1,\sigma,\lambda)$ for some $\sigma$ and $\lambda$ values.

The main properties of the distribution of $T$ are the following:

1. $aT \sim \text{GG}(a\eta,\sigma,\lambda)$ for all $a > 0$.
2. $T^a \sim \text{GG}(\eta^a, |a|\sigma, \text{sign}(a)\lambda)$ for all $a \neq 0$. For example, $T^{-1} \sim \text{GG}(\eta^{-1},\sigma,-\lambda)$.
3. $\text{E}(T^r) = \eta^r \dfrac{\Gamma(\lambda^{-2} + r\sigma\lambda^{-1})}{|\lambda|^{-2\frac{r\sigma}{\lambda}}\,\Gamma(\lambda^{-2})}$ for all $r$ such that $r\sigma\lambda > -1$. Therefore, when they exist, the coefficient of variation, the coefficient of skewness, and the coefficient of kurtosis of $T$ do not depend on $\eta$.
4. If $V \sim \text{G}(\alpha = \lambda^{-2}, \beta = \lambda^{-2})$, where $V$ has mean $\alpha\beta^{-1}$ and variance $\alpha\beta^{-2}$, then $T$ may be expressed as $\eta V^{\frac{\sigma}{\lambda}}$. Therefore,

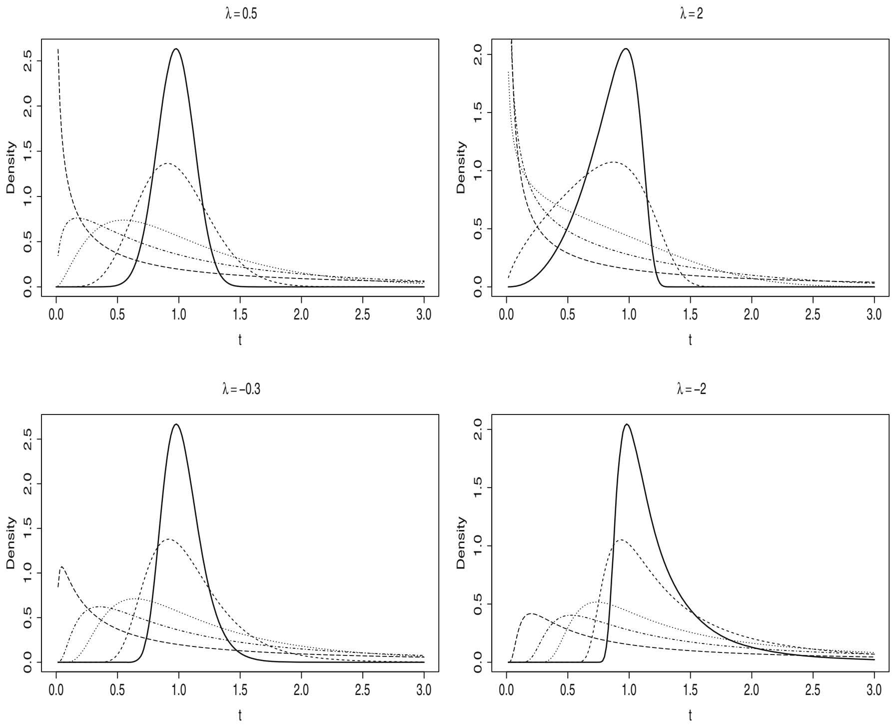


**Fig. 1** Graph of the probability density function of $T \sim \text{GG}(1, \sigma, \lambda)$ for some values of $\lambda$ and $\sigma = 0.15$ (——), $\sigma = 0.3$ (----), $\sigma = 0.7$ (·········), $\sigma = 1.1$ (— · —·) and $\sigma = 2$ (– – –)

(a) if $V_1, \ldots, V_n$ is a random sample drawn from $V \sim \text{G}(\lambda^{-2}, \lambda^{-2})$, then $\eta V_1^{\frac{\sigma}{\lambda}}, \ldots, \eta V_n^{\frac{\sigma}{\lambda}}$ is a random sample drawn from $T \sim \text{GG}(\eta, \sigma, \lambda)$.

(b) $f_T(t; \eta, \sigma, \lambda)$ may be expressed as $f_T(t; \eta, \sigma, \lambda) = f_V[(t/\eta)^{\frac{\lambda}{\sigma}}]\frac{|\lambda|}{\sigma\eta}(t/\eta)^{\frac{\lambda}{\sigma}-1}$, where $f_V(\cdot)$ is the p.d.f. of $V$.

(c) The c.d.f. (cumulative distribution function) of $T$, denoted as $F_T(\cdot)$, may be written as

$$F_T(t) = \begin{cases} F_V[(t/\eta)^{\frac{\lambda}{\sigma}}] & \text{for} \quad \lambda > 0 \\ 1 - F_V[(t/\eta)^{\frac{\lambda}{\sigma}}] & \text{for} \quad \lambda < 0, \end{cases}$$

where $F_V(\cdot)$ is the cumulative distribution function of $V$. The survival function of $T$ is given by $S_T(t) = 1 - F_T(t)$, so one may notice from the expression above that $\eta$ acts multiplicatively on the survival time $t$.

(d) The quantile function of $T$, denoted as $Q_T(\alpha)$, may be written as

$$Q_T(\alpha) = \begin{cases} \eta\big[Q_V(\alpha)\big]^{\frac{\sigma}{\lambda}} & \text{for} \quad \lambda > 0 \\ \eta\big[Q_V(1-\alpha)\big]^{\frac{\sigma}{\lambda}} & \text{for} \quad \lambda < 0, \end{cases}$$

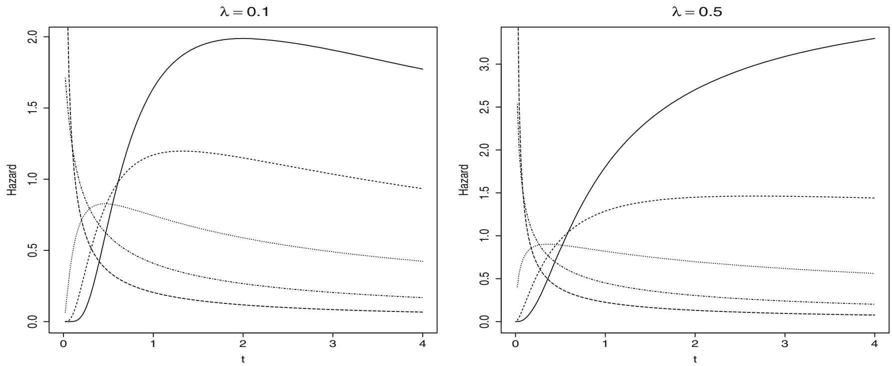


**Fig. 2** Graph of the hazard rate function of $T \sim \text{GG}(1, \sigma, \lambda)$ for some values of $\lambda$ and and $\sigma = 0.5$ (——), $\sigma = 0.7$ (----), $\sigma = 1.1$ (⋯⋯), $\sigma = 2$ (— · —·) and $\sigma = 4$ (– – –)

where $Q_V(\cdot)$ is the quantile function of $V$.

(e) The hazard rate function of $T$, denoted as $h_T(t)$, satisfies the following: $(i)$ If $\lambda = 1$ and $\sigma > 1$ or $1 < \lambda \le \sigma$ then $h_T(t)$ is decreasing; and $(ii)$ if $\lambda = 1$ and $\sigma < 1$ or $\sigma \le \lambda < 1$ then $h_T(t)$ is increasing.

(f) if $T_i = \text{GG}(\eta_i, \sigma_i, \lambda)$, for $i = 1, \ldots, n$, then $\sum_{i=1}^{n} (T_i/\eta_i)^{\frac{\lambda}{\sigma_i}} \sim \text{G}(n/\lambda^2, \lambda^{-2})$.

Figure 2 describes the hazard rate function of $T \sim \text{GG}(1, \sigma, \lambda)$ for some $\sigma$ and $\lambda$ values. One may have, for instance, constant, increasing, decreasing, bathtub-shaped and upside-down bathtub-shaped forms, among others, for the hazard rate function.

The parametric generalized gamma AFT model is defined as follows

1. $T_i = \eta_i\, \xi_i^{\sigma}$,
2. $\log(\eta_i) = \mathbf{x}_i^{\top}\boldsymbol{\beta}$ and
3. $\xi_i \overset{\text{iid}}{\sim} \text{GG}(1, 1, \lambda)$,

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\top}$ contains values of covariates and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\top}$ are the regression coefficients, $i = 1, \ldots, n$. Then, it follows that $T_i \overset{\text{ind}}{\sim} \text{GG}(\eta_i, \sigma, \lambda)$, and consequently the mean of $T_i$ is proportional to $\eta_i$ whereas the coefficient of variation and the shape of the hazard rate function of $T_i$ are constant for all observations. An appealing property of the parametric AFT model is that it allows to write the $100(\alpha)\%$ quantile of $T_i$ as follows

$$\log\big\{Q_{T_i}(\alpha)\big\} = \log\Big\{\eta_i\ Q_{\xi}^{\sigma}(\alpha)\Big\} = \sigma \log\big\{Q_{\xi}(\alpha)\big\} + \mathbf{x}_i^{\top}\boldsymbol{\beta},$$

where $Q_{\xi}(\alpha)$ is the $100(\alpha)\%$ quantile of $\xi$. Therefore, the model parameters also may be interpreted by taking into account their multiplicative effects acting on $Q_{T_i}(\alpha)$ for all $\alpha \in (0, 1)$. Application of generalized gamma models to study the survival of patients with AIDS under different therapies are given, for instance, in Cox et al. (2007).

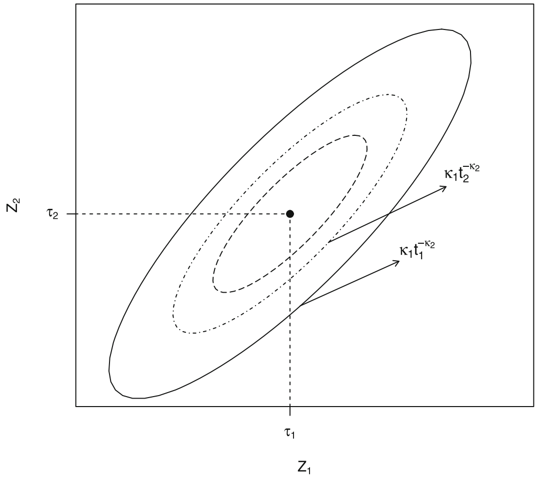


**Fig. 3** Description of the ellipses concerning the probabilities $S_T(t_1)$ and $S_T(t_2)$ for $t_2 > t_1$

### 2.2 The GG distribution as a lifetime model

Let suppose the status or condition of an object or individual may be quantified by comparing its values on a set of latent markers, denoted as $\mathbf{Z} = (Z_1, \ldots, Z_m)^\top$, with the "optimum"values of those markers, denoted as $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)^\top$. So, the probability that the lifetime of that object or individual exceeds a threshold, say $t$, decreases as the distance between $\mathbf{Z}$ and its "optimum"value $\boldsymbol{\tau}$ increases. That is,

$$S_T(t) = \mathrm{P}[T > t] = \mathrm{P}\left[(\mathbf{Z} - \boldsymbol{\tau})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{\tau}) \leq \kappa_1\, t^{-\kappa_2}\right],$$

for some $\kappa_1 > 0$ and $\kappa_2 > 0$, where $\boldsymbol{\tau}$ and $\boldsymbol{\Sigma}$ are the location and scale parameters of the distribution of $\mathbf{Z}$, respectively. $\mathbf{Z}$ may be assumed to have a very flexible $m$-dimensional distribution such as the multivariate power exponential distribution, studied by Gómez et al. (1998), with $\boldsymbol{\tau}$, $\boldsymbol{\Sigma}$ and $\nu$ parameters, where $\boldsymbol{\tau} \in \mathbb{R}^m$, $\boldsymbol{\Sigma}$ is an $m \times m$ definite positive symmetric matrix, and $\nu \in (0, \infty)$. This family of distributions includes as special cases the multivariate normal ($\nu = 1$), the multivariate Laplace or double exponential ($\nu = \frac{1}{2}$), and the multivariate uniform ($\nu \to \infty$). Then, the distribution of the lifetime of the object or individual, $T$, is

$$\mathrm{GG}\left(\left(\frac{\kappa_1 \nu}{m}\right)^{\frac{1}{\kappa_2}}, \frac{\sqrt{2\nu}}{\kappa_2\sqrt{m}}, -\frac{\sqrt{2\nu}}{\sqrt{m}}\right),$$

as, according to Gómez et al. (1998, p. 4), the distribution of $(\mathbf{Z} - \boldsymbol{\tau})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{\tau})$ is $\mathrm{G}(m/2\nu, 1/2)$ (see, illustration in Fig. 3 for $m = 2$).

Similarly, the status or condition of an object or individual may be quantified by comparing its values on a set of latent markers with the "worst"values of those markers. So, the probability that the lifetime of that object or individual exceeds a threshold, say $t$, increases as the distance between $\mathbf{Z}$ and its "worst"value $\boldsymbol{\tau}$ increases.

# 3 Generalized log-gamma additive partial linear models

## 3.1 Generalized log-gamma distribution

In particular $Y = \log(T)$ follows the generalized log-gamma distribution (see, for instance, Lawless 1980), namely GLG($\mu, \sigma, \lambda$), whose p.d.f. takes the form

$$f_Y(y; \mu, \sigma, \lambda) = \begin{cases} \frac{C(\lambda)}{\sigma} e^{\frac{1}{\lambda\sigma}(y-\mu) - \frac{1}{\lambda^2} e^{\frac{\lambda}{\sigma}(y-\mu)}} & \text{if } \lambda \neq 0 \\ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} & \text{if } \lambda = 0, \end{cases}$$

where $y \in \mathbb{R}$ and $\mu = \log(\eta)$. The location, scale and shape parameters are $\mu \in \mathbb{R}$, $\sigma > 0$ and $\lambda \in \mathbb{R}$, respectively. When $\lambda = 0$ one has the univariate normal family, if $\lambda = 1$ one has the extreme value distribution type I for the minimum and when $\lambda = -1$ the extreme value distribution type I for the maximum. Various other distributions may be derived as the logarithm transformation of the generalized gamma distribution. Another feature of the GLG family is that one has asymmetric distribution to the right if $\lambda < 0$ and asymmetric distribution to the left if $\lambda > 0$. Figure 4 describes the p.d.f. of the GLG($0, \sigma, \lambda$) for some $\sigma$ and $\lambda$ values. If $\lambda \neq 0$, the mean and variance of $Y$

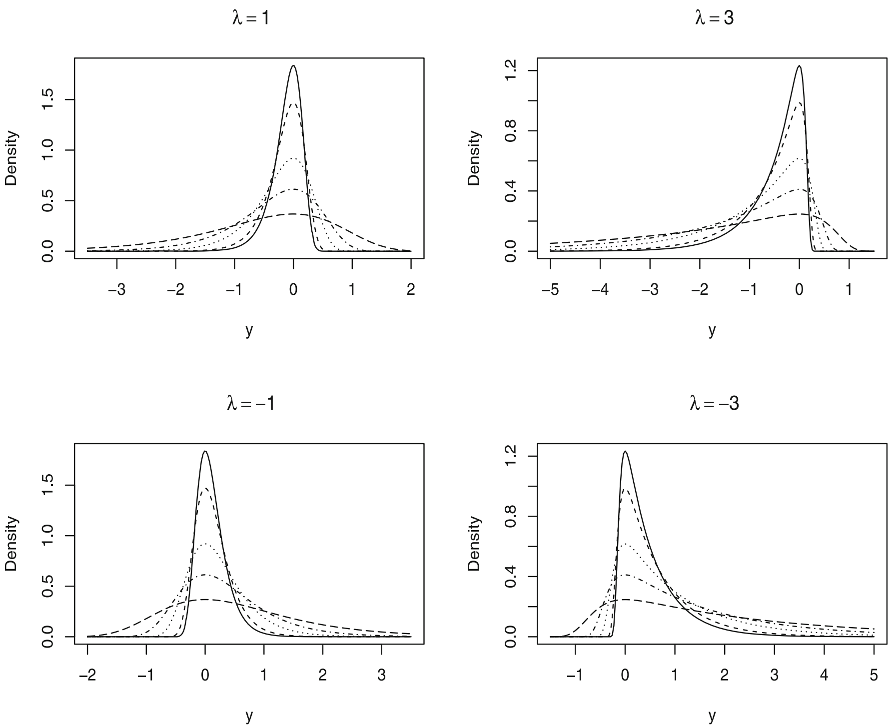


**Fig. 4** Graph of the probability density function of $Y \sim$ GLG($0, \sigma, \lambda$) for some values of $\lambda$ and and $\sigma = 0.2$ (——), $\sigma = 0.25$ (----), $\sigma = 0.4$ (·········), $\sigma = 0.6$ (— · —·) and $\sigma = 1$ (– – –)

are, respectively, given by

$$\mathrm{E}(Y) = \mu + \sigma \left\{ \frac{\psi(\lambda^{-2}) - \log(\lambda^{-2})}{|\lambda|} \right\} \text{ and } \mathrm{Var}(Y) = \sigma^2 \frac{\psi'(\lambda^{-2})}{\lambda^2},$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ are, respectively, the digamma and trigamma functions. A detailed study and on this family of distributions may be found, for instance, in Lawless (2003).

### 3.2 Additive partial linear models

A well known property of AFT models is that the covariates should act multiplicatively on the time $t$, as showed in Sect. 2.1 for generalized gamma AFT models. In the context of regression, to obtain directly this effect, the location parameter should be modeled on covariate values remaining constant the parameters $\sigma$ and $\lambda$. So, in this section, we will propose a semiparametric generalized log-gamma regression model in which the location parameter is modeled in an additive partial linear form with the scale and shape parameters being constant for all observations.

Then, let $y_1, \ldots, y_n$ be a set of random variable values and assume the following relationship between the response and covariate values:

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + f_1(t_{i1}) + \cdots + f_r(t_{ir}) + \sigma \epsilon_i, \tag{1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ denotes a $p \times 1$ vector of covariate values with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ being a $p \times 1$ vector of parametric regression coefficients, $f_\ell(t)$ are smoothing functions of the observed values $t_{i\ell}$ of additional continuous covariates, for example covariates of controlling, and $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{GLG}(0, 1, \lambda)$, for $i = 1, \ldots, n$ and $\ell = 1, \ldots, r$.

So, $Y_i$ has the distribution

$$Y_i \overset{\text{ind}}{\sim} \mathrm{GLG}\Big(\mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{\ell=1}^{r} f_\ell(t_{i\ell}), \sigma, \lambda\Big),$$

which implies that

$$\mathrm{E}(Y_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{\ell=1}^{r} f_\ell(t_{i\ell}) + \sigma \left\{ \frac{\psi(\lambda^{-2}) - \log(\lambda^{-2})}{|\lambda|} \right\} \text{ and } \mathrm{Var}(Y_i) = \sigma^2 \frac{\psi'(\lambda^{-2})}{\lambda^2},$$

for $i = 1, \ldots, n$. Model (1) may also be named generalized gamma AFT additive partial linear model.

### 3.3 B-splines

We will suppose the smoothing functions $f_\ell(t)$ approximated by B-splines (de Boor 1978 and Wood 2017, Chap. 5), namely $f_\ell(t) = \sum_{j=1}^{q_\ell} \mathrm{N}_{j\ell,k}(t)\gamma_{j\ell}$, where

$$\mathrm{N}_{j\ell,0}(t) = \begin{cases} 1 & t^0_{j\ell} \leq t < t^0_{(j+1)\ell} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathrm{N}_{j\ell,k}(t) = \omega_{j\ell,k}(t)\mathrm{N}_{j\ell,(k-1)}(t) + \{1 - \omega_{(j+1)\ell,k}(t)\}\mathrm{N}_{(j+1)\ell,(k-1)}(t),$$

$\omega_{j\ell,k}(t) = (t - t^0_{j\ell})/(t^0_{(j+k)\ell} - t^0_{j\ell})$, with $\mathrm{N}_{j\ell,k}(t)$ denoting the B-spline basis functions of degree $k$ and $\gamma_{j\ell}$ are coefficients to be estimated, whereas $m_\ell = q_\ell + k + 1$ denotes the number of internal knots, namely $a_\ell < t^0_{1\ell} < \cdots < t^0_{m_\ell \ell} < b_\ell$, for $j = 1, \ldots, q_\ell$, $\ell = 1, \ldots, r$ and $k = 1, 2, 3, \ldots$. We will assume in this paper $k = 3$ (cubic B-splines) so that the smoothing function $f(t)$ will be re-expressed as $f_\ell(t) = \sum_{j=1}^{q_\ell} \mathrm{N}_{j\ell}(t)\gamma_{j\ell}$ with $\mathrm{N}_{j\ell}(t) = \mathrm{N}_{j\ell,3}(t)$.

Therefore, (1) may be written as a linear model

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{N}_{i1}^\top \boldsymbol{\gamma}_1 + \cdots + \mathbf{N}_{ir}^\top \boldsymbol{\gamma}_r + \sigma\epsilon_i, \tag{2}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is a $p \times 1$ vector of explanatory variable values, $\mathbf{N}_{i\ell} = (\mathrm{N}_{1\ell}(t_{i\ell}), \ldots, \mathrm{N}_{q_\ell \ell}(t_{i\ell}))^\top$ is a $q_\ell \times 1$ vector with the $\ell$th cubic B-spline basis values, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $\boldsymbol{\gamma}_\ell = (\gamma_{1\ell}, \ldots, \gamma_{q_\ell \ell})^\top$ are the regression coefficients and $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{GLG}(0, 1, \lambda)$, for $i = 1, \ldots, n$ and $\ell = 1, \ldots, r$.

### 3.4 P-spline smoothing

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \sigma, \lambda)^\top$ with $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_r^\top)^\top$ the parameter vector to be estimated and the regular log-likelihood function expressed as

$$\mathrm{L}(\boldsymbol{\theta}) = n \log\left\{\frac{C(\lambda)}{\sigma}\right\} + \frac{1}{\lambda}\sum_{i=1}^{n} \epsilon_i - \frac{1}{\lambda^2}\sum_{i=1}^{n} e^{\lambda\epsilon_i},$$

where $\epsilon_i = (y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{N}_{i1}^\top \boldsymbol{\gamma}_1 - \cdots - \mathbf{N}_{ir}^\top \boldsymbol{\gamma}_r)/\sigma$ with $\mathbf{x}_i^\top$ and $\mathbf{N}_{i\ell}^\top$ being, respectively, the $i$th rows of $\mathbf{X}$ and $\mathbf{N}_\ell$, for $i = 1, \ldots, n$ and $\ell = 1, \ldots, r$. Since the number of parameters to be estimated increases with $s = q_1 + \cdots + q_r$ the direct maximization of $\mathrm{L}(\boldsymbol{\theta})$ may cause overfitting and non identification of $\boldsymbol{\gamma}$, so some penalization should be imposed for maximizing $\mathrm{L}(\boldsymbol{\theta})$. A usual procedure is to consider as penalty the integrated of the square of the second derivative of $f_\ell(t)$, which leads to the penalized log-likelihood function

$$\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \mathrm{L}(\boldsymbol{\theta}) - \sum_{\ell=1}^{r} \frac{\alpha_\ell}{2} \int_{a_\ell}^{b_\ell} \{f_\ell''(t)\}^2 dt, \qquad (3)$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r)^\top$ is named smoothing parameter vector that is estimated separately. One has $\alpha_\ell > 0$ and $\alpha_\ell \to 0$ implies data interpolation, whereas $\alpha_\ell \to \infty$ leads to linear approximation for $f_\ell(t)$. From Wood (2017, Chap. 5) the penalization in (3) may be expressed as

$$\int_{a_\ell}^{b_\ell} \{f_\ell''(t)\}^2 dt = \boldsymbol{\gamma}_\ell^\top \mathbf{M}_\ell \boldsymbol{\gamma}_\ell,$$

where $\mathbf{M}_\ell$ is a $q_\ell \times q_\ell$ positive semidefinite penalty matrix. However, Eilers and Marx (1996) showed that the integrated of the square of the $d$th derivative of $f_\ell(t)$ is well approximated by a penalty on finite differences of the coefficients $\boldsymbol{\gamma}_\ell$ with much less effort, namely

$$\int_{a_\ell}^{b_\ell} \{f_\ell^{(d)}(t)\}^2 dt \cong \sum_{j=d+1}^{q_\ell} [\Delta^d \gamma_{j\ell}]^2 = \boldsymbol{\gamma}_\ell^\top \mathbf{D}_{d\ell}^\top \mathbf{D}_{d\ell} \boldsymbol{\gamma}_\ell,$$

where $\mathbf{D}_{d\ell}$ denotes the penalty difference matrix of order $d$. This approach is named P-spline smoothing. In particular, for $d = 2$ and $q_\ell = 3$, one has that

$$\Delta^2 \gamma_{j\ell} = \gamma_{j\ell} - 2\gamma_{(j-1)\ell} + \gamma_{(j-2)\ell}$$

and

$$\mathbf{D}_{2\ell} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix}.$$

We will consider in this paper cubic B-splines and the approximation above for the integrated of the second derivative of each $f_\ell(t)$, for $\ell = 1, \ldots, r$. Nevertheless, extensions for other degrees of B-splines and penalty differences is straightforward requiring few changes in the notation.

Thus, the penalized log-likelihood function will be expressed as

$$\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \mathrm{L}(\boldsymbol{\theta}) - \sum_{\ell=1}^{r} \frac{\alpha_\ell}{2} \boldsymbol{\gamma}_\ell^\top \mathbf{M}_\ell \boldsymbol{\gamma}_\ell, \qquad (4)$$

where $\mathbf{M}_\ell = \mathbf{D}_{2\ell}^\top \mathbf{D}_{2\ell}$.The maximum likelihood penalized estimate (MLPE) of $\boldsymbol{\theta}$ will be obtained by maximizing (4) for $\boldsymbol{\alpha}$ fixed.

## 4 Parameter estimation

To perform an iterative procedure for the estimation of the MPLE of $\boldsymbol{\theta} = (\boldsymbol{\xi}^\top, \sigma, \lambda)^\top$, where $\boldsymbol{\xi} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$, the penalized score function and the penalized Fisher information matrix are derived, see details in Sects. S.1, S.2 and S.3.

### 4.1 Penalized score function and penalized Fisher information matrix

Using the results given in Sect. S.1 we express the penalized score function for $\boldsymbol{\theta} = (\boldsymbol{\xi}^\top, \sigma, \lambda)^\top$ as

$$\mathbf{U}_p^{\theta} = \begin{bmatrix} \mathbf{U}_p^{\xi} \\ \mathbf{U}_p^{\sigma} \\ \mathbf{U}_p^{\lambda} \end{bmatrix} = \begin{bmatrix} \frac{1}{\lambda\sigma}\mathbf{N}^\top\{\mathbf{D}(\mathbf{d}) - \mathbf{I}_n\}\mathbf{1}_n - \mathbf{M}_\alpha\boldsymbol{\xi} \\ \frac{1}{\lambda\sigma}\mathbf{1}_n^\top\{\mathbf{D}(\mathbf{d} - \mathbf{I}_n)\}\boldsymbol{\epsilon} - \frac{n}{\sigma} \\ n\zeta_\lambda - \frac{1}{\lambda^2}\mathbf{1}_n^\top\boldsymbol{\epsilon} + \frac{2}{\lambda^3}\mathbf{1}_n^\top\mathbf{D}(\mathbf{d})\mathbf{1}_n - \frac{1}{\lambda^2}\mathbf{1}_n^\top\mathbf{D}(\mathbf{d})\boldsymbol{\epsilon} \end{bmatrix},$$

where $\mathbf{N} = [\mathbf{X}, \mathbf{N}_1, \ldots, \mathbf{N}_r]$, $\mathbf{M}_\alpha = \text{blockdiag}\{\mathbf{O}_{pp}, \alpha_1\mathbf{M}_1, \ldots, \alpha_r\mathbf{M}_r\}$, $\mathbf{O}_{pp}$ is a $p \times p$ matrix of zeros, $\mathbf{I}_n$ denotes the identity matrix of order $n$, $\mathbf{1}_n$ is an $n \times 1$ vector of ones, $\mathbf{D}(\mathbf{d})$ is a diagonal matrix of $\mathbf{d} = (d_1, \ldots, d_n)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ with $d_i = e^{\lambda\epsilon_i}$ and $\epsilon_i = (y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - \mathbf{N}_{i1}^\top\boldsymbol{\gamma}_1 - \cdots - \mathbf{N}_{ir}^\top\boldsymbol{\gamma}_r)/\sigma$, for $i = 1, \ldots, n$, and $\zeta_\lambda = \frac{1}{\lambda} + \frac{2}{\lambda^3}\{\psi(\lambda^{-2}) + 2\log(|\lambda|) - 1\}$.

From Sects. S.2 and S.3 the penalized Fisher information matrix for $\boldsymbol{\theta}$ may be expressed as

$$\mathbf{K}_p^{\theta\theta} = \begin{bmatrix} \mathbf{K}_p^{\xi\xi} & \mathbf{K}_p^{\xi\sigma} & \mathbf{K}_p^{\xi\lambda} \\ \mathbf{K}_p^{\sigma\xi} & \mathbf{K}_p^{\sigma\sigma} & \mathbf{K}_p^{\sigma\lambda} \\ \mathbf{K}_p^{\lambda\xi} & \mathbf{K}_p^{\lambda\sigma} & \mathbf{K}_p^{\lambda\lambda} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\{\mathbf{N}^\top\mathbf{N} + \sigma^2\mathbf{M}_\alpha\} & \frac{u_\lambda}{\sigma^2}\mathbf{N}^\top\mathbf{1}_n & -\frac{u_\lambda}{\lambda\sigma}\mathbf{N}^\top\mathbf{1}_n \\ \frac{u_\lambda}{\sigma^2}\mathbf{1}_n^\top\mathbf{N} & \frac{n}{\sigma^2}(1 + v_\lambda) & \frac{n}{\lambda^2\sigma}\kappa_{2,\lambda} \\ -\frac{u_\lambda}{\lambda\sigma}\mathbf{1}_n^\top\mathbf{N} & \frac{n}{\lambda^2\sigma}\kappa_{2,\lambda} & \frac{n}{\lambda^2}\kappa_{1,\lambda} \end{bmatrix},$$

with

$$\begin{aligned}
\kappa_{1,\lambda} &= \tau_\lambda - \frac{2}{\lambda^2}\{\psi(\lambda^{-2}) - 2\log(|\lambda|)\} + \frac{6}{\lambda^2} - \frac{4}{\lambda}u_\lambda + v_\lambda, \\
\kappa_{2,\lambda} &= u_\lambda - \lambda v_\lambda - \frac{1}{\lambda}\{\psi(\lambda^{-2}) - 2\log(|\lambda|)\}, \\
\tau_\lambda &= 1 - \frac{1}{\lambda^2}\left\{10 - 12\log(|\lambda|) - 6\psi(\lambda^{-2}) - \frac{4}{\lambda^2}\psi'(\lambda^{-2})\right\}, \\
u_\lambda &= \mathrm{E}(\epsilon e^{\lambda\epsilon}) \\
&= \frac{1}{|\lambda|}\{\psi(1 + \lambda^{-2}) + 2\log(|\lambda|)\}, \\
v_\lambda &= \mathrm{E}(\epsilon^2 e^{\lambda\epsilon}) \\
&= \frac{1}{\lambda^2}[\psi'(1 + \lambda^{-2}) + \{\psi(1 + \lambda^{-2}) + 2\log(|\lambda|)\}^2]
\end{aligned}$$

and $\epsilon \sim \text{GLG}(0, 1, \lambda)$. Note that the parameters in $\boldsymbol{\theta} = (\boldsymbol{\xi}^\top, \sigma, \lambda)^\top$ are not orthogonal. In particular, for $\lambda = 1$, the distribution of $\epsilon$ is the standard extreme value distribution of type I for the minimum. Since $\psi(1) = -\gamma_m$, $\psi(2) = 1 - \gamma_m$, $\psi'(1) = \frac{\pi^2}{6}$ and $\psi'(2) = \frac{\pi^2}{6} - 1$, one has some simplifications, $\tau_1 = \frac{4\pi^2}{6} - 6\gamma_m - 9$, $u_1 = 1 - \gamma_m$ and $v_1 = \frac{\pi^2}{6} - 1 + (1 - \gamma_m)^2$, where $\gamma_m$ denotes de Euler–Mascheroni constant. Consequently, one obtains $\kappa_{1,1} = \frac{5\pi^2}{6} + (1-\gamma_m)^2 - 8$ and $\kappa_{2,1} = -\frac{\pi^2}{6} - (1-\gamma_m)^2 + 2$. On the other hand, for $\lambda = -1$, the distribution of $\epsilon$ is the standard extreme value distribution of type I for the maximum. Here, one has $\tau_{-1} = \tau_1$, $u_{-1} = u_1$ and $v_{-1} = v_1$, and then $\kappa_{1,-1} = \frac{5\pi^2}{6} + (1-\gamma_m)^2 - 8\gamma_m$ and $\kappa_{2,-1} = \frac{\pi^2}{6} + (1-\gamma_m)^2 - 2\gamma_m$.

### 4.2 Iterative processes

Different iterative processes may be proposed to obtain the MPLE $\hat{\boldsymbol{\theta}}$. For example, the well known backfitting (Gauss–Seidel) algorithm described, for instance, by Hastie and Tibshirani (1990, Chap. 6) and Green and Silverman (1994, Chap. 4). Although elegant and easy to be implemented this algorithm gets slow as the number of nonparametric components increases. On the other hand, penalized score Fisher algorithms may be applied to speed the convergence for obtaining $\hat{\boldsymbol{\theta}}$, such as extensions of the P-GAM algorithm proposed by Marx and Eilers (1996) for generalized additive models.

Then, consider the starting values $(\boldsymbol{\xi}^{(0)\top}, \sigma^{(0)}, \lambda^{(0)})^\top$ and that $\mathbf{N}$ is a full column rank matrix. The penalized Fisher scoring algorithm for obtaining the MPLE $\widehat{\boldsymbol{\xi}}$, given $\sigma$ and $\lambda$ and by remaining $\boldsymbol{\alpha}$ fixed, may be expressed as

$$\boldsymbol{\xi}^{(l+1)} = (\mathbf{N}^\top\mathbf{N} + \sigma^2\mathbf{M}_\alpha)^{-1}\mathbf{N}^\top\mathbf{y}_d^{(l)}, \tag{5}$$

for $l = 0, 1, \ldots$, where $\mathbf{y}_d = \boldsymbol{\mu} + \frac{\sigma}{\lambda}\{\mathbf{D}(\mathbf{d}) - \mathbf{I}_n\}\mathbf{1}_n$ is the pseudo-response and $\boldsymbol{\mu} = \mathbf{N}\boldsymbol{\xi}$. The iterative process (5) should be alternated with the following quasi-Newton iterative process for obtaining the MPLEs of $\sigma$ and $\lambda$:

$$(\sigma^{(m+1)}, \lambda^{(m+1)})^\top = \text{argmax}_{(\sigma,\lambda)}\mathrm{L}_p(\boldsymbol{\xi}^{(m+1)}, \sigma, \lambda, \boldsymbol{\alpha}), \tag{6}$$

for $m = 0, 1, 2\ldots$, where $\boldsymbol{\xi}^{(m+1)}$ corresponds to the profiled MPLEs of $\boldsymbol{\xi}$, obtained at the $(m + 1)$th convergence of the iterative process (5).

Then, for $\boldsymbol{\alpha}$ fixed, one has the following algorithm:

1. Give starting values for $\boldsymbol{\theta}$, keeping fixed $\sigma$ and $\lambda$ and performing the iterative process (5) for obtaining the profiled MPLE of $\boldsymbol{\xi}$. The starting values $\boldsymbol{\xi}^{(0)}$ and $\sigma^{(0)}$ are obtained from the unpenalized fit of the linear model (2) by keeping $\lambda$ fixed. It is assumed either $\lambda^{(0)} = -1$ or $\lambda^{(0)} = 1$ if there is asymmetry of the empirical distribution of $Y$ to right or to left, respectively.
2. Given the profiled MPLEs of $\boldsymbol{\xi}$ obtained at the $(m+1)$th convergence of the iterative process (5), obtain from (6) the profiled MPLEs of $\sigma$ and $\lambda$.
3. Alternating the iterative processes (5) and (6) until the joint convergence for obtaining the MPLE of $\boldsymbol{\theta}$ for $\boldsymbol{\alpha}$ fixed.

However, since one has $\mathbf{U}_p^{\theta}$ and $\mathbf{K}_p^{\theta\theta}$ written in closed-form expressions, one may apply, alternatively, a full penalized Fisher scoring algorithm for obtaining $\hat{\boldsymbol{\theta}}$, given by

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \{(\mathbf{K}_p^{\theta\theta})^{(m)}\}^{-1}(\mathbf{U}_p^{\theta})^{(m)},$$

for $m = 0, 1, 2\ldots$ and $\boldsymbol{\alpha}$ fixed. This iterative process requires fewer steps than the iterative process (5)–(6), so may speed the convergence, but simplifications such as in (5) are more difficult to be obtained.

For $\lambda = 0$ one has the normal case, whose MPLEs for $\alpha$ fixed are given by

$$\hat{\boldsymbol{\xi}} = (\mathbf{N}^{\top}\mathbf{N} + \widehat{\sigma}^2\mathbf{M}_{\alpha})^{-1}\mathbf{N}^{\top}\mathbf{y}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{N}\hat{\boldsymbol{\xi}})^{\top}(\mathbf{y} - \mathbf{N}\hat{\boldsymbol{\xi}}).$$

For a grid of $\boldsymbol{\alpha}$ values, the MPLE of $\boldsymbol{\theta}$ is obtained by minimizing either the Akaike (1974) criterion or the Schwarz (1978) criterion, respectively, defined as

$$\mathrm{AIC}(\boldsymbol{\alpha}) = -2\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\alpha}) + 2\{2 + \mathrm{df}(\boldsymbol{\alpha})\}$$

and

$$\mathrm{BIC}(\boldsymbol{\alpha}) = -2\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\alpha}) + \log(n)\{2 + \mathrm{df}(\boldsymbol{\alpha})\},$$

where $\mathrm{df}(\boldsymbol{\alpha})$ denotes the effective degrees of freedom that will be defined in Sect. 6. Alternatively, the generalized cross-validation method (see, for instance, Wood 2017, Chap. 6), which consists in minimizing the function,

$$\mathrm{GCV}(\alpha) = \frac{n\sum_{i=1}^{n}\{y_i - \widehat{\mathrm{E}}(Y_i)\}^2}{\{n - 2 - \mathrm{df}(\boldsymbol{\alpha})\}^2}$$

may be applied for selecting an appropriate smoothing parameter. A faster option to obtain $\boldsymbol{\alpha}$ is a direct maximization of the function

$$\mathrm{L}_p(\boldsymbol{\xi}^{(m+1)}, \sigma^{(m+1)}, \lambda^{m+1)}, \boldsymbol{\alpha})$$

after each cycle of the algorithm (1)–(3). For example, by using the procedure `optim` available in the `R` package. The Gauss-Seidel and Fisher scoring algorithms for maximizing (4) are implemented in the `R` package `sglg` (Cardozo et al. 2021) available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=sglg.

## 5 Effective degrees of freedom

The main idea behind the definition of effective degrees of freedom is trying to provide the cost of estimating the parameters in the linear predictor $\boldsymbol{\mu} = \mathbf{N}\boldsymbol{\xi}$, for $\boldsymbol{\alpha}$ fixed. A useful procedure consists in defining an appropriate smoother that projects a pseudo-response onto $\hat{\boldsymbol{\mu}}$ and then to take the sum of its eigenvalues as the effective degrees of freedom, denoted by $\text{df}(\boldsymbol{\alpha})$ (see, for instance, Hastie and Tibshirani 1990, Chap. 5).

In order to derive $\text{df}(\boldsymbol{\alpha})$ for the MPLE $\hat{\boldsymbol{\xi}}$ we will consider the following expression of the iterative process (5) at the convergence:

$$\hat{\boldsymbol{\xi}} = \{\mathbf{N}^{\top}\mathbf{N} + \hat{\sigma}^2\mathbf{M}_{\alpha}\}^{-1}\mathbf{N}^{\top}\hat{\mathbf{z}}_d,$$

where $\hat{\mathbf{z}}_d = \hat{\boldsymbol{\eta}} + \frac{\hat{\sigma}}{\hat{\lambda}}\{\mathbf{D}(\hat{\mathbf{d}}) - \mathbf{I}_n\}\mathbf{1}_n$. It follows that $\hat{\boldsymbol{\mu}} = \mathbf{N}\hat{\boldsymbol{\xi}} = \widehat{\mathbf{H}}(\boldsymbol{\alpha})\hat{\mathbf{z}}_d$ with

$$\widehat{\mathbf{H}}(\boldsymbol{\alpha}) = \mathbf{N}\{\mathbf{N}^{\top}\mathbf{N} + \hat{\sigma}^2\mathbf{M}_{\boldsymbol{\alpha}}\}^{-1}\mathbf{N}^{\top}$$

being named smoother or linear predictor for $\boldsymbol{\alpha}$ fixed. The effective degrees of freedom is defined as $\text{df}(\boldsymbol{\alpha}) = \text{tr}\{\widehat{\mathbf{H}}(\boldsymbol{\alpha})\}$. From Eilers and Marx (1996) (see also Vanegas and Paula 2016) one has that

$$\begin{aligned}
\text{df}(\boldsymbol{\alpha}) &= \text{tr}\{\mathbf{N}\{\mathbf{N}^{\top}\mathbf{N} + \hat{\sigma}^2\mathbf{M}_{\alpha}\}^{-1}\mathbf{N}^{\top}\} \\
&= \text{tr}\{\{\mathbf{N}^{\top}\mathbf{N} + \hat{\sigma}^2\mathbf{M}_{\alpha}\}^{-1}\mathbf{N}^{\top}\mathbf{N}\} \\
&= \text{tr}\left\{\left(\mathbf{I}_{p+s} + \hat{\sigma}^2\mathbf{Q}^{-\frac{1}{2}}\mathbf{M}_{\alpha}\mathbf{Q}^{-\frac{1}{2}}\right)^{-1}\right\} \\
&= \sum_{i=1}^{p+s}\frac{1}{1+\phi_i(\boldsymbol{\alpha})},
\end{aligned}$$

where $\phi_i(\boldsymbol{\alpha}) \geq 0$ are the eigenvalues of the non-negative definite matrix $\hat{\sigma}^2\mathbf{Q}^{-\frac{1}{2}}\mathbf{M}_{\alpha}\mathbf{Q}^{-\frac{1}{2}}$ with $\mathbf{Q}^{\frac{1}{2}}$ being a positive definite matrix such that $\mathbf{Q}^{\frac{1}{2}}\mathbf{Q}^{\frac{1}{2}} = \mathbf{N}^{\top}\mathbf{N}$. Since the first $p$ eigenvalues of $\hat{\sigma}^2\mathbf{Q}^{-\frac{1}{2}}\mathbf{M}_{\alpha}\mathbf{Q}^{-\frac{1}{2}}$ are zeros it follows that

$$\text{df}(\boldsymbol{\alpha}) = p + \sum_{i=p+1}^{p+s}\frac{1}{1+\phi_i(\boldsymbol{\alpha})}.$$

Additionally, one has $p + s > \text{df}(\boldsymbol{\alpha}) > p +$ zero, where zero denotes the number of zeros eigenvalues of $\mathbf{M}_{\alpha}$. The effective degrees of freedom associated to $\hat{\boldsymbol{\gamma}}_{\ell}$, namely $\text{df}(\alpha_{\ell})$, correspond to the sum of the principal diagonal elements of $\widehat{\mathbf{H}}(\boldsymbol{\alpha})$ from the $(p+q_1+\cdots+q_{(\ell-1)})$th position to the $(p+q_1+\cdots+q_{\ell})$th position, for $\ell = 1, \ldots, r$. Then, one has that $\text{df}(\boldsymbol{\alpha}) = p + \sum_{\ell=1}^{r}\text{df}(\alpha_{\ell})$.

### 5.1 Inference

The inference for $\boldsymbol{\theta}$ may be based on the assumption of asymptotic normality for $\hat{\boldsymbol{\theta}}$ with the approximate variance-covariance matrix obtained from the inverse of the penalized Fisher information matrix $\mathbf{K}_p^{\theta\theta}$. This result has support on the Bayesian approach for the model $\mathbf{y} = \mathbf{N}\boldsymbol{\xi} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ follows a multivariate normal distribution. Then, by assuming an improper prior for $\boldsymbol{\xi}$ the posterior distribution $\boldsymbol{\xi}|\mathbf{y}$ follows a multivariate normal distribution of mean $\hat{\boldsymbol{\xi}}$ and variance-covariance matrix that corresponds to the inverse of the respective penalized Fisher information matrix for $\boldsymbol{\xi}$. So, credible intervals may be constructed for any quantities derived from $\boldsymbol{\gamma}$. As pointed out by Wood (2017, Sect. 6.10) this result may be extended for large $n$ for regression models fitted by penalized iteratively re-weighted least square algorithms.

Another inference of interest is to construct asymptotic confidence bands for each function vector $\mathbf{f}_\ell = (f_\ell(t_{1\ell}), \ldots, f_\ell(t_{n\ell}))^\top$. Such bands, named pointwise confidence bands (see, for instance, Vanegas and Paula 2016), are formed by joining the asymptotic confidence intervals for $f_\ell(t_{1\ell}), \ldots, f_\ell(t_{n\ell})$ which are centered in $\hat{\mathbf{f}}_\ell$ and whose standard errors are derived from the principal diagonal of the approximate variance-covariance matrix $\mathrm{Var}(\hat{\mathbf{f}}_\ell) = \mathbf{N}_\ell \mathrm{Var}(\hat{\boldsymbol{\gamma}}_\ell)\mathbf{N}_\ell^\top$.

## 6 Diagnostic procedures

The aim of diagnostic procedures is to assess departures from the assumption made for the model, such as error assumptions and the functional form between the location parameter and the covariates, to detect outlying observations and to perform sensitivity studies of the parameter estimates under perturbations made in the model/data. In this section we will discuss the derivation of the quantile residual (Dunn and Smyth 1996) for model (1) and the curvatures of local influence (Cook 1986) to assess the sensitivity of the parameter estimates under small perturbation in the model/data.

### 6.1 Quantile residual

The aim of residual analysis is to detect outlying observations and to assess important departures from the assumptions made for the error distribution in regression models. The quantile residual proposed by Dunn and Smyth (1996) have been largely applied due to its easy interpretation and may be performed from the c.d.f. of the postulated error distribution. 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 $F_Y(y; \boldsymbol{\theta})$ and $\Phi(\cdot)$ denote, respectively, the c.d.f. of $Y$ and the c.d.f. of N(0, 1). For large $n$ and under the postulated model $r_{q_1}, \ldots, r_{q_n}$ follow i.i.d. standard normal distribution. So, the normal probability plot between $r_{q_i}$, for $i = 1, \ldots, n$, and the quantiles of the standard normal distribution may assess the goodness-of-fit of the fitted model.

The c.d.f. of $Y \sim \text{GLG}(\mu, \sigma, \lambda)$ is given by

$$F_Y(y; \boldsymbol{\theta}) = \begin{cases} 1 - \text{I}(\lambda^{-2}, e^{\lambda \frac{(y-\mu)}{\sigma}}) & \text{if } \lambda > 0 \\ \text{I}(\lambda^{-2}, e^{\lambda \frac{(y-\mu)}{\sigma}}) & \text{if } \lambda < 0 \\ \Phi\left(\frac{y-\mu}{\sigma}\right) & \text{if } \lambda = 0, \end{cases}$$

where $\text{I}(k, x) = \frac{1}{\Gamma(k)} \int_0^x \nu^{k-1} e^{-\nu} d\nu$ and $\Gamma(k)$ denote, respectively, the incomplete gamma function and the gamma function. The survival function and the hazard function of $Y$ are defined as $S_Y(y; \boldsymbol{\theta}) = 1 - F_Y(y; \boldsymbol{\theta})$ and $h_Y(y; \boldsymbol{\theta}) = f_Y(y; \boldsymbol{\theta})/S_Y(y; \boldsymbol{\theta})$, respectively.

### 6.2 Local influence

A key reason behind of diagnostic methods is to answer the following natural and important question: how the parameter estimates change under perturbations in the model/data? One quite original approach to this question was given by Cook (1986). Through the perturbed log-likelihood function Cook assessed the influence of small perturbations in the model/data on the parameter estimates. So, a reasonable extension to semiparametric models is

$$\text{L}_p(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{\omega}) = \text{L}(\boldsymbol{\theta}|\boldsymbol{\omega}) - \sum_{l=1}^{r} \frac{\alpha_l}{2} \boldsymbol{\gamma}_l^\top \mathbf{M}_l \boldsymbol{\gamma}_l,$$

which we call perturbed penalized log-likelihood function, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)^\top$ is a perturbation vector in some open subset $\Omega$ of $I\!R^m$ in which lives $\boldsymbol{\omega}_0$ a reference point in the analysis. At $\boldsymbol{\omega}_0$ the perturbed penalized log-likelihood function satisfies

$$\text{L}_p(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{\omega}_0) = \text{L}_p(\boldsymbol{\theta}, \boldsymbol{\alpha}).$$

So, we have the non-perturbed penalized log-likelihood function. Cook also employed the concept of the likelihood displacement function, given by

$$\text{LD}(\boldsymbol{\omega}) = 2[\text{L}_p(\hat{\boldsymbol{\theta}}, \boldsymbol{\alpha}|\boldsymbol{\omega}_0) - \text{L}_p(\hat{\boldsymbol{\theta}}_\omega, \boldsymbol{\alpha}|\boldsymbol{\omega})],$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_\omega$ are the maximum penalized likelihood estimates of $\boldsymbol{\theta}$ under $\text{L}_p(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{\omega}_0)$ and $\text{L}_p(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{\omega})$, respectively. Let $\phi(\boldsymbol{\omega}) = (\boldsymbol{\omega}^\top, \text{LD}(\boldsymbol{\omega}))^\top$ be the influence graph. If the likelihood displacement function is a smooth function then the influence graph will be a regular surface and we can define normal curvatures. The essential idea, in the Cook's approach, is to study the normal curvatures for $\phi(\cdot)$ at $(\hat{\boldsymbol{\theta}}, \boldsymbol{\omega}_0)$ in the unitary direction $\boldsymbol{\ell}$. Such curvature is expressed as

$$\text{C}_\ell(\boldsymbol{\theta}) = 2 \mid \boldsymbol{\ell}^\top \boldsymbol{\Delta}^\top (\mathbf{J}_p^{\hat{\theta}\hat{\theta}})^{-1} \boldsymbol{\Delta} \boldsymbol{\ell} \mid,$$

where $\mathbf{J}_p^{\theta\theta}$ denotes the observed information matrix of $\boldsymbol{\theta}$, $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_{\beta\omega}^\top, \boldsymbol{\Delta}_{\gamma\omega}^\top, \Delta_{\sigma\omega}^\top, \Delta_{\lambda\omega}^\top)^\top$ is a $(p+s+2) \times n$ matrix with $\boldsymbol{\Delta}_{\beta\omega}$ being a $p \times n$ matrix with elements $(\Delta_{\beta\omega})_{ji} = \partial \mathrm{L}_p^2(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{\omega})/\partial\beta_j\partial\omega_i$ for $j = 1, \ldots, p$ and $i = 1, \ldots, n$, whereas $\boldsymbol{\Delta}_{\gamma\omega} = (\boldsymbol{\Delta}_{\gamma_1\omega}^\top, \ldots, \boldsymbol{\Delta}_{\gamma_r\omega}^\top)^\top$ with $\boldsymbol{\Delta}_{\gamma_l\omega}$ being $q_l \times n$ matrices with elements $(\Delta_{\gamma_l\omega})_{jil} = \partial \mathrm{L}_p^2(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{\omega})/\partial\gamma_{jl}\partial\omega_i$ for $j = 1, \ldots, q_l$, $i = 1, \ldots, n$ and $l = 1, \ldots, r$. Similarly, $\boldsymbol{\Delta}_{\sigma\omega}$ and $\boldsymbol{\Delta}_{\lambda\omega}$ are $1 \times n$ vectors with elements $(\Delta_{\sigma\omega})_i = \partial \mathrm{L}_p^2(\boldsymbol{\theta}, \alpha|\boldsymbol{\omega})/\partial\sigma\partial\omega_i$ and $(\Delta_{\lambda\omega})_i = \partial \mathrm{L}_p^2(\boldsymbol{\theta}, \alpha|\boldsymbol{\omega})\partial\lambda\partial\omega_i$ for $i = 1, \ldots, n$, respectively. All the quantities are evaluated at $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega}_0$.

Cook suggests to take the index plot of $|\boldsymbol{\ell}_{max}|$, the largest eigenvector relative to the largest eigenvalue $\mathrm{C}_{max}$ of the matrix $\boldsymbol{\Delta}^\top(\mathbf{J}_p^{\hat{\theta}\hat{\theta}})^{-1}\boldsymbol{\Delta}$. The observations pointed out in the graph are suspected to have large influence on the parameter estimates under the adopted perturbation scheme. Alternatively, Poon and Poon (1999) proposed the conformal normal curvature in order to have a curvature invariant under uniform change of scale, defined as

$$\mathrm{B}_\ell(\boldsymbol{\theta}) = \frac{|\,\boldsymbol{\ell}^\top\boldsymbol{\Delta}^\top(\mathbf{J}_p^{\hat{\theta}\hat{\theta}})^{-1}\boldsymbol{\Delta}\boldsymbol{\ell}\,|}{\sqrt{\mathrm{tr}\Big(\{\boldsymbol{\Delta}^\top(\mathbf{J}_p^{\hat{\theta}\hat{\theta}})^{-1}\boldsymbol{\Delta}\}^2\Big)}}.$$

This curvature has the property that $0 \leq \mathrm{B}_\ell(\boldsymbol{\theta}) \leq 1$ for any unitary direction $\boldsymbol{\ell}$ and an aggregate influential measure for all $q$-influential eigenvectors is defined as

$$m(q)_i = \sqrt{\sum_{j=1}^{k} \hat{\lambda}_i \ell_{ji}^2},$$

where $\hat{\lambda}_{max} = \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_k \geq q/\sqrt{n} > \hat{\lambda}_{k+1}, \ldots, \hat{\lambda}_n \geq 0$ denote the ordered normalized eigenvalues of $\boldsymbol{\Delta}^\top(\mathbf{J}_p^{\hat{\theta}\hat{\theta}})^{-1}\boldsymbol{\Delta}$ and $\ell_{ji}$ is the $i$th element of the $j$th unitary eigenvector $\boldsymbol{\ell}_j$, 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 assess those observations that are influential for all eigenvectors such that $\mathrm{B}_{\ell_j}(\boldsymbol{\theta}) \geq q/\sqrt{n}$. A particular influential measure that considers the conformal normal curvature in the direction of the $i$th observation is defined as

$$\mathrm{B}_{e_i}(\boldsymbol{\theta}) = \mathrm{B}_i = m^2(0)_i = \sum_{i=1}^{n} \hat{\lambda}_j \ell_{ji}^2,$$

where $\mathbf{e}_i$ is an $n \times 1$ vector of zeros with 1 at the $i$th position. The measure $\mathrm{B}_i$ may be interpreted as the square of the total contribution of the unitary eigenvectors. According to Lee and Xu (2004) we may use $\mathrm{B}_i > \bar{\mathrm{B}} + c^*\mathrm{SD(B)}$ to discriminate if an observation is suspect or not to be influential, where $\bar{\mathrm{B}}$ and SD(B) denote, respectively, the mean and the standard deviation of $\mathrm{B}_i$, for $i = 1, \ldots, n$ and $c^*$ is selected appropriately.

In particular, if one has the partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$ and the interest is on the subvector $\boldsymbol{\theta}_1$, for example $\boldsymbol{\theta}_1 = \boldsymbol{\beta}$, $\boldsymbol{\theta}_1 = \boldsymbol{\gamma}$, $\theta_1 = \sigma$ or $\theta_1 = \lambda$, one may perform the index plot of the largest eigenvector relative to the largest eigenvalue $C_{max}$ of the matrix $\boldsymbol{\Delta}^\top \mathbf{B}_1 \boldsymbol{\Delta}$ or then to evaluate the conformal normal curvature

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

where $\mathbf{B}_1 = (\mathbf{J}_p^{\hat{\theta}\hat{\theta}})^{-1} - \mathrm{blockdiag}\{\mathbf{0}, (-\mathbf{J}_p^{\hat{\theta}_2\hat{\theta}_2})^{-1}\}$ and $\mathbf{J}_p^{\theta_2\theta_2}$ is the observed information matrix of $\boldsymbol{\theta}_2$. Here the index plot of $\mathrm{B}_{e_i}(\boldsymbol{\theta}_1)$ may also be considered.

#### 6.2.1 Perturbation schemes

**6.2.1.1 Case-weight perturbation** For this perturbation scheme we assume that

$$\mathrm{L}_p(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \mathrm{L}_i(\boldsymbol{\theta}) - \sum_{l=1}^{r} \frac{\alpha_l}{2} \boldsymbol{\gamma}_l^\top \mathbf{M}_l \boldsymbol{\gamma}_l,$$

where $\mathrm{L}_i(\boldsymbol{\theta}) = \log\left\{\frac{C(\lambda)}{\sigma}\right\} + \frac{\epsilon_i}{\lambda} - \frac{e^{\lambda\epsilon_i}}{\lambda^2}$ with $\epsilon_i = (y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{N}_1^\top \boldsymbol{\gamma}_1 - \cdots - \mathbf{N}_r^\top \boldsymbol{\gamma}_r)/\sigma$, $0 \le \omega_i \le 1$, for $i = 1, \ldots, n$, and $\boldsymbol{\omega}_0 = (1, \ldots, 1)^\top$ is the reference point.

Then, we obtain

$$(\Delta_{\beta\omega})_{ji}\Big|_{\omega_0} = -\frac{x_{ij}}{\lambda\sigma} + \frac{x_{ij} e^{\lambda\epsilon_i}}{\lambda\sigma}, \quad (\Delta_{\gamma_{l\omega}})_{kil}\Big|_{\omega_0} = -\frac{\mathrm{N}_{kil}}{\lambda\sigma} + \frac{\mathrm{N}_{kil} e^{\lambda\epsilon_i}}{\lambda\sigma},$$
$$(\Delta_{\sigma\omega})_i\Big|_{\omega_0} = -\frac{1}{\sigma} - \frac{\epsilon_i}{\lambda\sigma} + \frac{\epsilon_i e^{\lambda\epsilon_i}}{\lambda\sigma} \text{ and } (\Delta_{\lambda\omega})_i\Big|_{\omega_0} = \zeta_\lambda - \frac{\epsilon_i}{\lambda^2} + \frac{2e^{\lambda\epsilon_i}}{\lambda^3} - \frac{\epsilon_i e^{\lambda\epsilon_i}}{\lambda^2},$$

where $\zeta_\lambda = \frac{1}{\lambda} + \frac{2}{\lambda^3}\{\psi(\lambda^{-2}) + 2\log(|\lambda|) - 1\}$, $\mathrm{N}_{kil} = \mathrm{N}_{kl}(t_{i\ell})$, for $i = 1, \ldots, n$, $j = 1, \ldots, p$, $k = 1, \ldots, q_l$ and $l = 1, \ldots, r$, with all the parameters being evaluated at the estimate $\widehat{\boldsymbol{\theta}}$.

**6.2.1.2 Response perturbation** For this perturbation scheme we will consider that each $y_i$ is perturbed as

$$y_{i_\omega} = y_i + \omega_i.$$

where $\omega_i \in I\!R$, for $i = 1, \ldots, n$. Hence $\epsilon_{i_\omega} = (y_{i_\omega} - \mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{N}_1^\top \boldsymbol{\gamma}_1 - \cdots - \mathbf{N}_r^\top \boldsymbol{\gamma}_r)/\sigma$ with $\boldsymbol{\omega}_0 = (0, \ldots, 0)^\top$. So, we obtain

$$(\Delta_{\boldsymbol{\beta}\omega})_{ji}\Big|_{\omega_0} = \frac{x_{ij} e^{\lambda\epsilon_i}}{\sigma^2}, \quad (\Delta_{\gamma_l\omega})_{kil}\Big|_{\omega_0} = \frac{\mathrm{N}_{kil} e^{\lambda\epsilon_i}}{\sigma^2},$$
$$(\Delta_{\sigma\omega})_i\Big|_{\omega_0} = \frac{1}{\lambda\sigma^2}\Big(\lambda\epsilon_i e^{\lambda\epsilon_i} + e^{\lambda\epsilon_i} - 1\Big) \text{ and } (\Delta_{\lambda\omega})_i\Big|_{\omega_0} = \frac{1}{\lambda\sigma}\Big(\frac{e^{\lambda\epsilon_i}}{\lambda} - \epsilon_i e^{\lambda\epsilon_i} - \frac{1}{\lambda}\Big),$$

for $i = 1, \ldots, n$, $j = 1, \ldots, p$, $k = 1, \ldots, q_l$ and $l = 1, \ldots, r$, with all the parameters being evaluated at the estimate $\hat{\boldsymbol{\theta}}$.

## 7 Simulation studies

In this section we will describe the results from simulation studies performed to assess the empirical behavior of the MPLEs from the partially linear model

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + f(t_i) + \sigma\epsilon_i, \tag{7}$$

with $\sigma > 0$ and $\epsilon_i \sim \text{GLG}(0, 1, \lambda)$, $f(t_i) = \cos(4\pi t_i)$ for $t_i = (2i-1)/2n$ and $i = 1, \ldots, n$. The values assigned for the parameters are $\beta_1 = 2$, $\beta_2 = 0.5$, $\sigma = 1.0$ and 2, and $\lambda = -1, -0.6, 0.4$ and 1. The explanatory variable values $x_{i1}$'s were generated from Bernoulli with probability 0.5, whereas $x_{i2}$'s were generated from a uniform distribution in the interval [−3,3]. These variable values were generated and fixed for each sample size of $n = 50, 100, 200$ and 500. We considered 10 internal knots equally spaced and the value $\alpha = 0.025$ for the smoothing parameter.

In order to study the consistency of the MPLEs $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}$ and $\hat{\lambda}$ and their asymptotic normality we computed the mean squared error (MSE), the bias (in absolute value) and the statistic of the Kolmogorov-Smirnov (KS) goodness-of-fit test for each estimator under each scenario and sample size. The MSE and bias for the parameter $\theta$ were calculated, respectively, as $R^{-1}\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \theta_0)^2$ and $|\bar{\hat{\theta}} - \theta_0|$, where $\bar{\hat{\theta}} = R^{-1}\sum_{r=1}^{R}\hat{\theta}^{(r)}$ with $\hat{\theta}^{(r)}$ being the estimate of $\theta$ from the $r$th replicate and $\theta_0$ is the true parameter value. The consistency of the cubic B-spline estimator $\mathbf{N}\widehat{\boldsymbol{\gamma}}$ is verified by graphing the shape of the true function $f(t) = \cos(4\pi t)$ jointly with the average estimates $\mathbf{N}\bar{\hat{\boldsymbol{\gamma}}}$ for each scenario and sample size considered, where $\bar{\hat{\boldsymbol{\gamma}}} = R^{-1}\sum_{r=1}^{R}\hat{\boldsymbol{\gamma}}^{(r)}$ with $\hat{\boldsymbol{\gamma}}^{(r)}$ being the estimate of $\boldsymbol{\gamma}$ from the $r$th replication. We considered for each scenario $R = 5000$ replicates. The algorithm of this simulation study is described below:

1. Fix the number of replications $R$, the sample size $n$, the true parameter vector $\boldsymbol{\theta}_0 = (\beta_{01}, \beta_{02}, \sigma_0, \lambda_0)^\top$, the nonlinear component $f(t)$ and $\alpha$. Then, generate the variable values of $X_1 \sim \text{Be}(0.5)$ and $X_2 \sim \text{U}[-3, 3]$.
2. Generate $R$ random samples of size $n$ of the response variable $Y$, based on the equation (7).
3. Fit $M_1, M_2, \ldots, M_R$ generalized log-gamma partially linear models from the algorithm given in Sect. 4.2.
4. Then, from the MPLEs $\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(R)}$ calculate the MSE, bias and KS statistic for the respective parameter setting and sample size.

The results are summarized in Tables S1-S4 and Figures S1-S8. We may notice from Tables S1-S4 indication of consistency and asymptotic normality of the four estimators considered and that the bias and MSE in general decreases as the sample size increases. However, the consistency of the dispersion parameter estimator seems slower than the consistency of the remaining parameter estimators. Looking at Figures S1-S8, one may observe strong indication that the cubic B-spline estimator converge to the true function $f(t)$.

## 8 Application

In order to illustrate the methodologies developed in this paper we will analyze a subset of the data set on 769 personal injury insurance claims made in Australia from January 1998 to June 1999 and described in de Jong and Heller (2008, pp. 4 and 14–15). Claims settled with zero payment were not included. In this data set the following variables were considered: `amount` (amount of paid money by an insurance policy or claim size, in Australian dollars), `legrep` (with legal representation or not), `month` (month of occurrence of the accident) and `optime` (operational time in % with range [0, 100]). Operation time is known in the insurance area as the percentage of cases settle faster in the group than the given case. So, it is expected longer delays for larger claim sizes. Since we are analyzing only the years of 1998 and 1999 of the insurance data, the operation time for these cases changes in the range [0.1; 31.9]. Similarly to Paula et al. (2012), that analyzed the group with legal representation by applying log-Birnbaum-Saunders-t error models, we will also discard the predictor `month` due to the lack of correlation with the response variable `log(amount)`. However, our analysis we be concentrated on the group without legal representation. Thus, 227 personal injury insurance claims will be analyzed in this study.

Figure 5 describes the empirical density (left) and the robust boxplot (right) (Hubert and Vandervieren 2008) of log(amount) and one may notice a left-skewed distribution indicating that a log-gamma distribution with positive shape parameter may be suitable to describe the distribution of log(amount). The robust boxplot points out four observations with the largest values for log(amount) as possible outliers. In addition, the scatter plot between log(amount) and optime, described in Fig. 6, presents a nonlinear tendency suggesting a nonlinear model to explain the location of log(amount) given the optime.

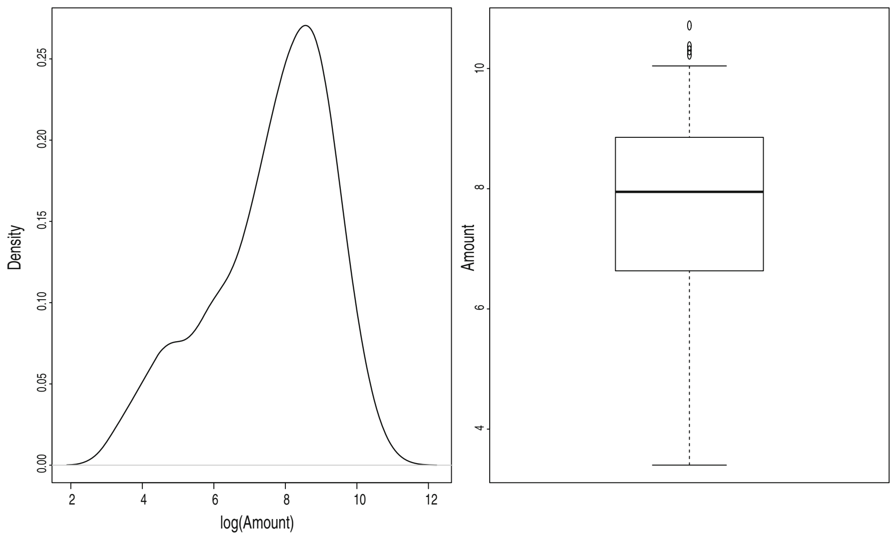


**Fig. 5** Density estimate (left) and robust boxplot (right) of the response log(amount)

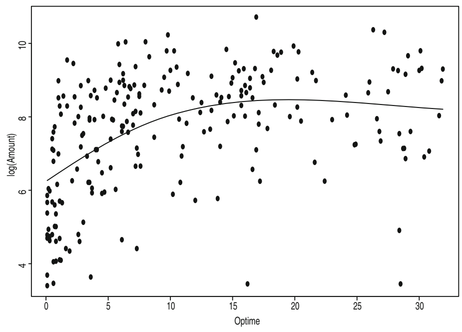


**Fig. 6** Scatter plot between log(amount) and optime with the tendency obtained from the command `smooth.spline` available in the R software

**Table 1** Parameter estimates and their approximate standard errors for the intercept, scale and shape parameters, and cubic B-spline coefficients as well as the smoothing parameter estimate, basis dimension and effective degrees of freedom from the fit of the partially linear model (8) to the insurance data

| Parameter | Estimate | SE | z value |
|---|---|---|---|
| $\mu$ | −1.1718 | 2.5735 | 0.65 |
| $\sigma$ | 1.2143 | 0.0687 | 17.67 |
| $\lambda$ | 0.7152 | 0.1620 | 4.42 |
| $\gamma_1$ | 8.4069 | 2.9179 | 2.88 |
| $\gamma_2$ | 9.3179 | 2.5856 | 3.60 |
| $\gamma_3$ | 9.8456 | 2.5380 | 3.88 |
| $\gamma_4$ | 9.8598 | 2.5736 | 3.83 |
| $\gamma_5$ | 9.9353 | 2.5795 | 3.85 |
| $\gamma_6$ | 9.7850 | 2.5748 | 3.80 |
| $\gamma_7$ | 9.9288 | 2.6155 | 3.80 |
| $\gamma_8$ | 10.6258 | 2.7871 | 3.81 |
| Covariate | $\alpha$ | $q$ | df($\alpha$) |
| Optime | 2 | 8 | 4.36 |

Then, according to the above descriptive analysis we will propose to fit this data set the following partially linear model:

$$Y_i = \mu + f(t_i) + \sigma\epsilon_i, \tag{8}$$

where $Y_i = \log(\text{amount}_i)$, $t_i = \text{optime}_i$ and $f(t) = \sum_{j=1}^{q} \mathrm{N}_{j,3}(t)\gamma_j$, where $\mathrm{N}_{j,3}(t)$ is expressed recursively

$$\mathrm{N}_{j,3}(t) = \omega_{j,3}(t)\mathrm{N}_{j,2}(t) + \{1 - \omega_{(j+1),3}(t)\}\mathrm{N}_{(j+1),2}(t),$$

$\omega_{j,3}(t) = (t - t_j^0)/(t_{(j+1)}^0 - t_j^0)$, with $\mathrm{N}_{j,2}(t)$ being computed from $\mathrm{N}_{j,1}(t)$ and $\mathrm{N}_{(j+1),1}(t)$, whereas $\mathrm{N}_{j,1}(t)$ is computed from $\mathrm{N}_{j,0}(t)$ and $\mathrm{N}_{(j+1),0}(t)$, for $t_j^0 \le t < t_{(j+1)}^0$, $j = 1, \ldots, q$, and $\epsilon_i \sim \mathrm{GLG}(0, 1, \lambda)$, $i = 1, \ldots, 227$.

The number of knots was selected from a rough manner by changing $q$ in the interval $[4, 8]$ until to guarantee a good fit. We choose $q = 8$ and consequently $m = 12$ internal knots $(t_1^0, \ldots, t_{12}^0)$ calculated from the sample quantiles of optime. The parameter estimates of the selected model are presented in Table 1. Note that the

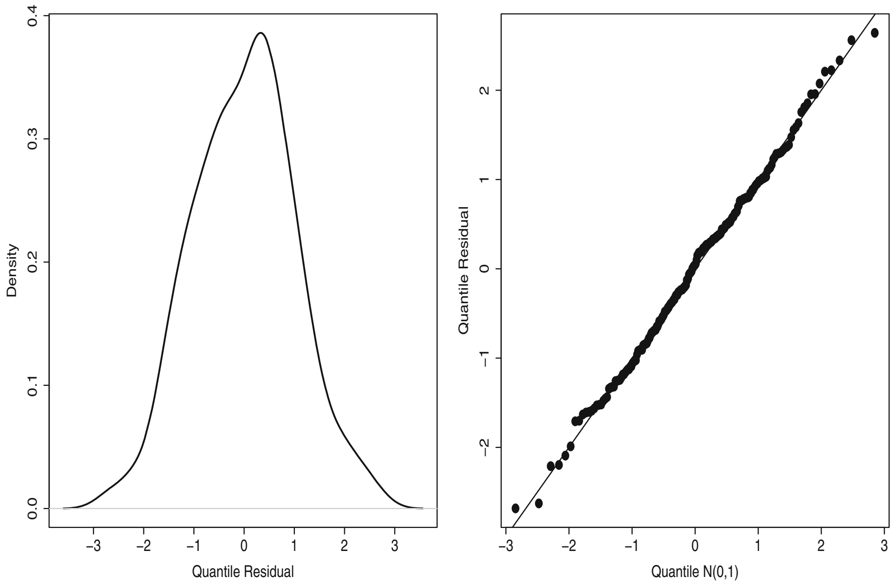


**Fig. 7** Empirical distribution of the quantile residual (left) and normal probability plot of the quantile residual(right) from the fit of the partially linear model (8) to the insurance data

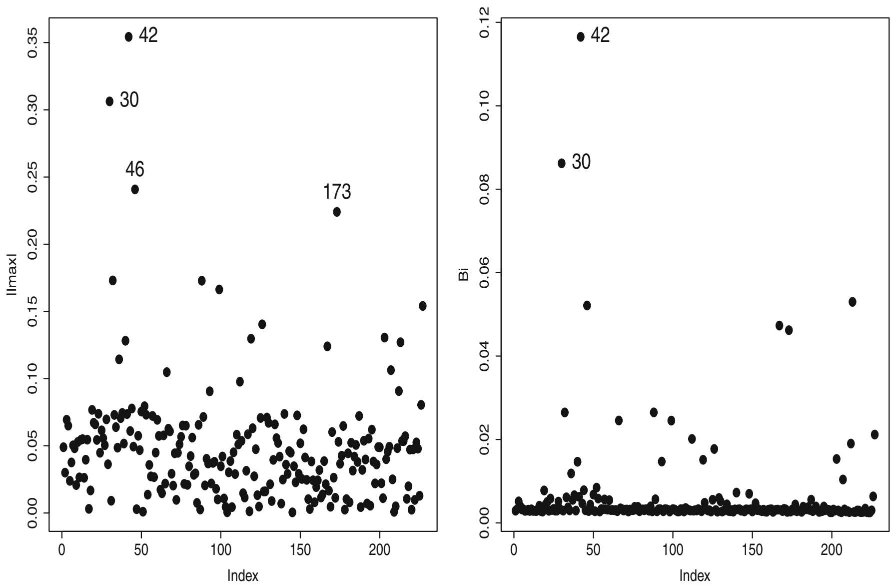


**Fig. 8** Index plots of $|\ell_{max}|$ (left) and $\mathrm{B}_i$ (right) under case-weight perturbation scheme from the fit of the partially linear model (8) to the insurance data

95% interval estimate for the shape parameter is given by [0.3977; 1.0327], which is according with the descriptive analysis. All the coefficients of the cubic B-spline appear marginally significant confirming the existence of a nonlinear tendency between the

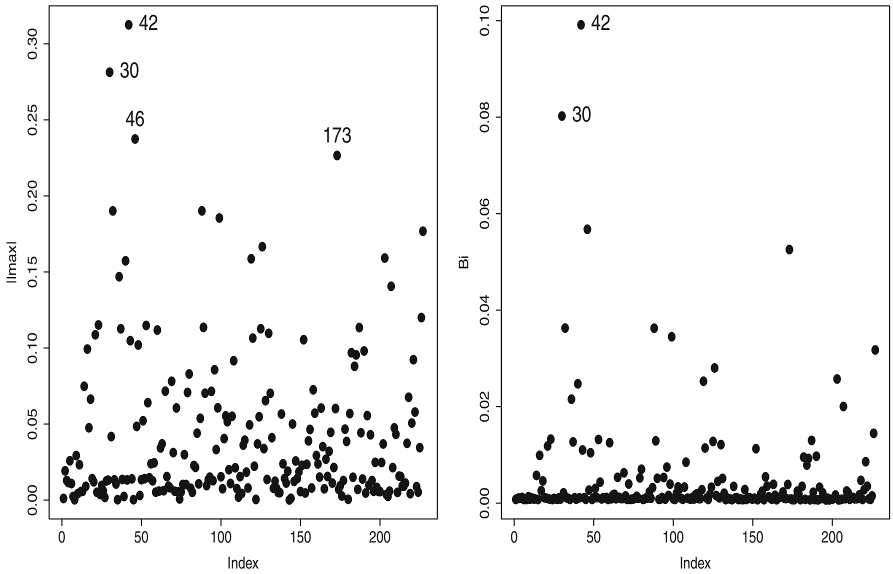


**Fig. 9** Index plots of $|\ell_{max}|$ (left) and $\mathrm{B}_i$ (right) under response perturbation scheme from the fit of the partially linear model (8) to the insurance data

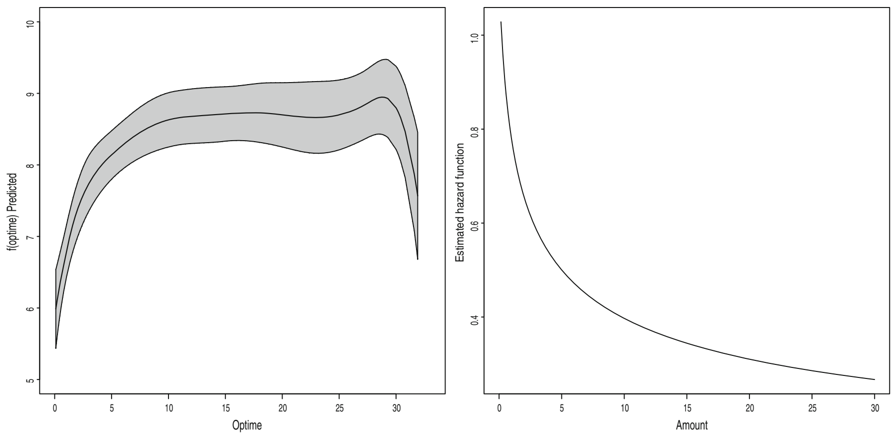


**Fig. 10** Pointwise confidence band for $f$(optime) (left) and the estimated hazard rate function from the fit of the partially linear model (8) to the insurance data

location of log(amount) and optime. However, since one has $\mathrm{df}(\alpha) = 4.36$ degrees of freedom associated with the MPLE $\hat{\xi} = (\hat{\mu}, \hat{\boldsymbol{\gamma}}^\top)^\top$, is expected smaller p-values for the estimated coefficients $\hat{\gamma}_1, \ldots, \hat{\gamma}_8$ than the ones based on the standard normal distribution.

From Fig. 7, that presents the empirical distribution and the normal probability plot (left) of the quantile residual (right), one may observe indication of a suitable fit. Figures 8 and 9 present the index plots of $|\ell_{max}|$ and $\mathrm{B}_i$ under the case-weight and

**Table 2** Results of fitting partially linear models under some asymmetric error distributions to the insurance data

| Distribution | #Par | AIC | Distribution | #Par | AIC |
|---|---|---|---|---|---|
| Generalized log-gamma | 3 | 785.43 | Skew-t type 1 | 4 | 789.78 |
| Skew-normal type 1 | 3 | 796.53 | Skew-t type 2 | 4 | 792.78 |
| Skew-normal type 2 | 3 | 783.04 | Skew-t type 3 | 4 | 784.83 |
| Skew-PE type 1 | 4 | 789.04 | Skew-t type 4 | 4 | 789.41 |
| Skew-PE type 2 | 4 | 795.27 | Generalized B2 | 4 | 4243.10 |

response perturbation schemes, respectively. Some observations are pointed out as possible influential on the parameter estimates. Particularly, the values of log(amount) and optime for the observations #30, #42 and #46 disagree with the tendency observed in Fig. 6. Their large values for log(amount), \$8.98, \$9.55 and \$9.45, correspond to small values for optime, 1, 1.7 and 2.2, respectively. In the sample the median values of log(amount) and optime are \$7.95 and 7.50, respectively. The other observation #173 pointed out in the graphs has the largest amount \$10.72 and a large optime 16.9. Finally, in Fig. 10 (left) one has the pointwise confidence band of 95% for the nonlinear function $f$(optime) and from Fig. 10 (right) one may notice that the estimated hazard rate function decreases as the claim size increases. An interpretation for this graph is that the probability of the insurer to pay a determinate amount for the insured, given that the insured paid for this amount in the insurance policy, decreases as the amount increases.

Table 2 summarizes the AIC from the partially linear model (8), under 9 different skew distributions available in GAMLSS (Righy et al. 2020), fitted to the insurance data. We may notice that the partially linear model (8) under the generalized log-gamma distribution presents an excellent performance taking into account the number of parameters and the qqplot of the quantile residual (see a comparison among the qqplots in Figures S9–S18). In addition, the generalized log-gamma error model has interesting features pointed out in Sects. 2 and 3, such as the equivalence with the GG AFT partially linear model and the possibility of performing quantile regression directly from the location parameter regression. Also, a P-GAM type iterative process is easily developed due to the closed-form expression derived for the Fisher information matrix.

## 9 Conclusions

Generalized gamma AFT partially linear models under uncensored data and P-spline smoothing are proposed in this paper. A penalized Fisher scoring iterative process is developed for estimating the regression coefficients of the parametric and nonparametric components, which is alternated with a quasi-Newton algorithm for estimating the dispersion and shape parameters. The effective degrees of freedom are derived from the proposed iterative process. Extensions of the quantile residual and the local influence approach are also performed as well as a simulation studies to assess the consistency and asymptotic normality of the parametric and nonparametric estimators. An illustrative example is given, in which the amount of paid money by an insurance policy in Australia is fitted by a nonlinear model given the operational time (in %) and for the group without legal representation. All the numerical analyzes and graphs were performed in the R package sglg (Cardozo et al. 2021) available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=sglg. The codes developed for the simulation studies and application are presented in Sect. S.6.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00362-022-01300-4.

**Acknowledgements** The authors are grateful to the associate editor and two reviewers for their helpful comments and suggestions. The first and second authors were partially supported by the Brazilian source CNPq, whereas the first author was also supported by the Colombian source COLCIENCIAS.

## References

Akaike H (1974) A new look at the statistical identification model. IEEE Trans Autom Control 19:716–723
Agostinelli C, Marazzi A, Yohai V (2014) Robust estimators of the generalized log-gamma distribution. Technometrics 56:92–101
Agostinelli C, Marazzi A, Yohai V (2017) robustloggamma: robust estimation of the generalized log gamma model. http://CRAN.R-project.org/package=robustloggamma
Amoroso L (1925) Richerche intorno alla curve die redditi. Ann Math Pure Appl 21:123–159
Cardozo CA, Paula GA, Vanegas L (2021) Fitting semi-parametric generalized log-gamma regression models. http://CRAN.R-project.org/package=sglg
Cook D (1986) Assessment of local influence (with discussion). J R Stat Soc B 48:133–169
Cox C, Chu H, Shneider MF, Munoz A (2007) Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distributions. Statistics in Medicine 26:4352–4374
de Boor C (1978) A practical guide for splines. Springer, New York
de Jong P, Heller GZ (2008) Generalized linear models for insurance data. Cambridge University Press, Cambridge
Dunn PK, Smyth GK (1996) Randomized quantile residuals. J Comput Graph Stat 5:236–244
Eilers P, Marx B (1996) Flexible smoothing with B-spline and penalties. Stat Sci 11:89–102
Fabio L, Paula GA, de Castro M (2012) A Poisson mixed model with nonnormal random effect distribution. Comput Stat Data Anal 56:1499–1510
Gómez E, Gómez-Villegas MA, Martín JM (1998) A multivariate generalization of the power exponential family of distributions. Commun Stat Theory Methods 27:589–600
Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models. Chapman and Hall/CRC, London
Hager H, Bain L (1970) Inferential procedures for the generalized gamma distribution. J Am Stat Assoc 65:1601–1609
Hastie T, Tibshirani R (1990) Generalized additive models. Chapman and Hall/CRC, London
Hubert M, Vandervieren E (2008) An adjusted boxplot for skewed distributions. Comput Stat Data Anal 52:5186–5201
Lawless JF (1980) Inference in the generalized gamma and log-gamma distributions. Technometrics 22:409–419
Lawless JF (2003) Statistical models and methods for life time data, 2nd edn. Wiley-Interscience, New York
Lee SY, Xu L (2004) Influence analyses of nonlinear mixed-effects models. Comput Stat Data Anal 45:321–341
Marx B, Eilers P (1996) Direct generalized additive modeling with penalized likelihood. Comput Stat Data Anal 28:193–209

Stasinopoulos MD, Righy RA, Gillian ZA, Voudouris V, de Bastiani F (2017) Flexible regression and smoothing using GAMLSS in R. Chapman and Hall/CRC, Boca Raton
Ortega E, Paula GA, Bolfarine H (2003) Influence diagnostics in generalized log-gamma regression models. Comput Stat Data Anal 42:165–186
Ortega E, Cancho V, Paula GA (2009) Generalized log-gamma regression models with cure fraction. Lifetime Data Anal 15:79–106
Paula GA, Rojas O (1997) On restricted hypothesis in extreme value regression models. Comput Stat Data Anal 25:143–157
Paula GA, Leiva V, Barros M, Liu S (2012) Robust statistical modeling using the Birnbaum–Saunders-t distribution applied to insurance. Appl Stoch Models Bus Ind 28:16–34
Poon W, Poon Y (1999) Conformal normal curvature and assessment of local influence. J R Stat Soc 61:51–61
Prentice R (1974) A log gamma model and its maximum likelihood estimation. Biometrika 61:539–542
R Core Team (2019) R: a language and environment for statistical computing. https://www.R-project.org/
Righy RA, Stasinopoulos MD, Gillian ZA, de Bastiani F (2020) Distributions for modeling location, scale, and shape using GAMLSS in R. Chapman and Hall/CRC, Boca Raton
Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6:461–464
Stacy E (1962) A generalization of the gamma distribution. Ann Math Stat 33:1187–1192
Stacy E, Mihram G (1965) Parameter estimation for a generalized gamma distribution. Technometrics 7:349–358
Vanegas L, Paula GA (2016) An extension of log-symmetric regression models: R codes and applications. J Stat Comput Simul 86:1709–1735
Wood S (2017) Generalized additive models: an introduction with R, 2nd edn. Chapman and Hall/CRC, London