Pak.j.stat.oper.res. Vol.16 No.4 2020 pp 827-838 DOI: http://dx.doi.org/10.18187/pjsor.v16i4.3443

Pakistan Journal of Statistics and Operation Research

# The normal-tangent-$G$ class of probabilistic distributions: properties and real data modelling

Fábio V. J. Silveira[1], Frank Gomes-Silva[2*], Cícero C. R. Brito[3], Moacyr Cunha-Filho[4], Jader S. Jale[5], Felipe Gusmão[6], Sílvio Xavier-Júnior[7]

*Corresponding author

1. Department of Statistics and Informatics, Rural Federal University of Pernambuco, Brazil, fabiovjs@gmail.com
2. Department of Statistics and Informatics, Rural Federal University of Pernambuco, Brazil, franksinatrags@gmail.com
3. Federal Institute of Education, Science and Technology of Pernambuco, Brazil, cicerocarlosbrito@yahoo.com.br
4. Department of Statistics and Informatics, Rural Federal University of Pernambuco, Brazil, moacyr2006@gmail.com
5. Department of Statistics and Informatics, Rural Federal University of Pernambuco, Brazil, jsjale1983@gmail.com
6. Department of Statistics and Informatics, Rural Federal University of Pernambuco, Brazil, felipe556@gmail.com
7. Department of Statistics, Paraíba State University, Brazil, silvioxj@gmail.com

**Abstract**

This paper introduces a novel class of probability distributions called normal-tangent-$G$, whose submodels are parsimonious and bring no additional parameters besides the baseline's. We demonstrate that these submodels are identifiable as long as the baseline is. We present some properties of the class, including the series representation of its probability density function (pdf) and two special cases. Monte Carlo simulations are carried out to study the behavior of the maximum likelihood estimates (MLEs) of the parameters for a particular submodel. We also perform an application of it to a real dataset to exemplify the modelling benefits of the class.

**Key Words:** Class of probabilistic distributions; Identifiable; Normal-tangent-$G$ class; Maximum likelihood; Special submodels; Inference.

## 1. Introduction

Works on new probability distributions are vast in the statistical literature; the authors generally have the intent to propose extended models that provide more flexibility than the commonly used ones do. For this purpose, many generalizations of probabilistic distributions are formulated by incorporating one or more additional parameters. To this extent, Mudholkar and Srivastava (1993) presented one of the most common ways to perform this technique, wherein they exponentiated the Weibull cumulative distribution function (cdf) to a shape parameter. Besides making the distribution more flexible, it allowed the hazard rate function (hrf) to have new shapes.

The procedure mentioned above was used by many authors to create generalized distributions. For instance, De Gusmão et al. (2011) exponentiated the cdf of the complementary Weibull model, discussed by Drapella (1993), to a shape parameter $\gamma > 0$ and created a three-parameter distribution called generalized inverse Weibull (GiW, for short), whose cdf is $F_{GiW}(x|\alpha,\beta,\gamma) = e^{-\gamma\alpha^{\beta}x^{-\beta}}$. The inclusion of $\gamma$ brings more flexibility; however, it is not seldom that distributions with many parameters may present major problems, like non-identifiability. Cordeiro and Castro (2011) intro-

duced a new family of generalized distributions called Kumaraswamy-$G$, whose cdf is $F_G(x) = 1 - [1 - G(x)^a]^b$, where $a > 0$, $b > 0$ and $G(x)$ is the cdf of a parent distribution. This generator is versatile and able to produce powerful submodels, whose extra parameters are related to the skewness and tail weights. Now, let us consider the Kumaraswamy-Exponential model where $\boldsymbol{\theta} = (\lambda, a, b)$ is the corresponding parametric vector; since $F_G(x|\boldsymbol{\theta}) = 1 - \left[1 - (1 - e^{-\lambda})^a\right]^b$, it is easy to see that if $\boldsymbol{\theta_1} = (\lambda_1, 1, b)$ and $\boldsymbol{\theta_2} = \left(\lambda_2, 1, \frac{\lambda_1}{\lambda_2} b\right)$, then $F_G(x|\boldsymbol{\theta}_1) = F_G(x|\boldsymbol{\theta}_2)$ for $\{\lambda_1, \lambda_2, b\} \subset \mathbb{R}_+^*$. In other words, there can be an infinite number of parametric vectors associated with the same cdf. Such property is undesirable and may entail complications regarding estimation of parameters. The GiW undergoes the same problem as well for any pair of parametric vectors $(\alpha_1, \beta, \gamma)$ and $\left(\alpha_2, \beta, \gamma \left[\frac{\alpha_1}{\alpha_2}\right]^\beta\right)$, where $\{\alpha_1, \alpha_2, \beta, \gamma\} \subset \mathbb{R}_+^*$.

The list of classes like the Kumaraswamy-$G$, that is, classes containing at least one unidentifiable submodel (even when the baseline $G$ is identifiable) is large. Some recent examples are the exponentiated logarithmic generated family (Marinho et al., 2018), the Gompertz-$G$ family (Alizadeh et al., 2017), the Lindley family (Cakmakyapan and Ozel, 2017) and the exponentiated Kumaraswamy-$G$ class (Silva et al., 2019). It does not mean that all of the submodels from the aforementioned classes are unidentifiable nor that one should keep from using them. Even so, researching and studying classes of probability distributions that somehow furnish identifiable submodels is worthwhile. For example, as long as the baseline $G$ is identifiable, any distribution emerged from the normal-$G$ class (Silveira et al., 2019) is identifiable; its submodels are generally more flexible than the original baselines, and they have the advantage of adding no extra parameters to the cdf. Some models, like the normal-Weibull, are able to fit asymmetrical data with either positive or negative skewness. Moreover, the distributions from the normal-$G$ class can be continuous or discrete, depending on the baseline. It was designed according to the method of generating classes of probability distributions presented by Brito et al. (2019).

The method has a high power of generalization. It introduces a general expression given by:

$$F(x) = \zeta(x) \sum_{j=1}^{n} \int_{L_j(x)}^{U_j(x)} \mathrm{d}H(t) - \nu(x) \sum_{j=1}^{n} \int_{M_j(x)}^{V_j(x)} \mathrm{d}H(t)\,, \tag{1}$$

where $H$ is a cdf, $n \in \mathbb{N}$, $\zeta, \nu : \mathbb{R} \mapsto \mathbb{R}$ and $L_j, U_j, M_j, V_j : \mathbb{R} \mapsto \mathbb{R} \cup \{\pm\infty\}$; certain conditions on these functions must be satisfied to assure that $F$ is a cdf. Setting $n = 1$, $\zeta(x) = 1$, $\nu(x) = 0$, $L_1(x) = -\infty$, $U_1(x) = \tan\left(\pi\left[G(x) - \frac{1}{2}\right]\right)$ and $H(t) = \Phi(t)$, where $\Phi$ is the standard normal cdf, this paper brings the normal-tangent-$G$ class of probability distributions (NT-$G$, for short), whose cdf is:

$$F_G(x) = \int_{-\infty}^{\tan\left(\pi\left[G(x) - \frac{1}{2}\right]\right)} \mathrm{d}\Phi(t)\,, \tag{2}$$

such that $G(x)$ is the baseline cdf. This novel class shares some important features with the normal-$G$ (particularly with regard to identifiability) but also presents new potentialities. We investigate some interesting models from the NT-$G$ class and make comparisons with other well-known extended probability distributions.

## 2. The normal-tangent-$G$ class and some mathematical properties

Given that $U_1(x) = \tan\left(\pi\left[G(x) - \frac{1}{2}\right]\right)$ and $U_1^\star(x) = \frac{2G(x)-1}{G(x)[1-G(x)]}$ are the upper limits of the integral in (1) for the NT-$G$ and the normal-$G$ classes respectively, it is easy to verify that $\lim_{x\to-\infty} U_1(x) = -\infty = \lim_{x\to-\infty} U_1^\star(x)$, $\lim_{x\to+\infty} U_1(x) = +\infty = \lim_{x\to+\infty} U_1^\star(x)$ and that $U_1(x)$ is a non-decreasing function. In agreement with the method (Brito et al., 2019), these conditions are enough to certify that (2) is a cdf, since the normal-$G$ cdf and (2) would be identical functions except for $U_1$ and $U_1^\star$.

Considering that $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, and $\Phi(x) = \int_{-\infty}^{x} \phi(t)\mathrm{d}t$, one can write the NT-$G$ cdf as follows:

$$F_G(x) = \Phi\left[\tan\left(\pi\left[G(x) - \frac{1}{2}\right]\right)\right]\,. \tag{3}$$

Notice that there are no additional parameters besides those of $G(x|\boldsymbol{\theta})$. In case of continuous $G(x)$, one can take the

derivative of (3) with respect to $x$ to obtain the following pdf:

$$f_G(x) = \phi\left[\tan\left(\pi\left[G(x) - \frac{1}{2}\right]\right)\right]\sec^2\left(\pi\left[G(x) - \frac{1}{2}\right]\right)\pi g(x)\,, \tag{4}$$

where $g(x)$ is the pdf of the random variable whose cdf is $G(x)$. The hrf of the NT-$G$ class is given by:

$$\tau_G(x) = \frac{\phi\left[\tan\left(\pi\left[G(x) - \frac{1}{2}\right]\right)\right]}{1 - \Phi\left[\tan\left(\pi\left[G(x) - \frac{1}{2}\right]\right)\right]}\sec^2\left(\pi\left[G(x) - \frac{1}{2}\right]\right)\pi g(x)\,. \tag{5}$$

The lack of identifiability is a problem that may cause some complications on parametric estimation since the uniqueness of the estimates is uncertain in this case. Hence, inferences on the parameters of an identifiable distribution are more reliable. As Theorem 2.1 states, members of the NT-$G$ class are always identifiable as long as the baseline $G$ is.

**Theorem 2.1.** *If the cdf $F_G$ belongs to the NT-$G$ class and the cdf $G$ is identifiable, then $F_G$ is identifiable.*

*Proof.* Given that $0 < G(x|\boldsymbol{\theta}_i) < 1$ for $i = 1, 2$, where $\boldsymbol{\theta}_i$ is a parametric vector, let us assume that $F_G(x|\boldsymbol{\theta}_1) = F_G(x|\boldsymbol{\theta}_2)$. Considering that the composed function $\Phi \circ \tan : \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \mapsto \mathbb{R}_+$ is injective, we can infer from (3) that $\pi\left[G(x|\boldsymbol{\theta}_1) - \frac{1}{2}\right] = \pi\left[G(x|\boldsymbol{\theta}_2) - \frac{1}{2}\right] \Rightarrow G(x|\boldsymbol{\theta}_1) = G(x|\boldsymbol{\theta}_2) \Rightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. □

### 2.1. Quantile Function

The quantile function $Q_F(p)$ of the NT-$G$ class is obtained without difficulty by inverting (3). That is:

$$Q_F(p) = Q_G\left(\frac{1}{\pi}\arctan\left[\Phi^{-1}(p)\right] + \frac{1}{2}\right) \tag{6}$$

where $Q_G$ is the quantile function of the parent distribution $G$. We use $Q_F$ along with a uniform random number generator to perform the simulation of random variables following (3). Namely, if $Z \sim \mathcal{U}(0, 1)$, then $Q_F(Z) \sim$ NT-$G$. An advantage of the NT-$G$ class over the normal-$G$ is that (6) is continuous on the real interval $(0, 1)$, whereas the normal-$G$ quantile function has a point of discontinuity at $p = 0.5$. Thus, it would not be an obstacle for the NT-$G$ class to define some meaningful quantities that depend on the median, like the Galton's skewness.

### 2.2. Shapes

The characterization of the distribution shape and the number of modes can be determined by examining the critical points of the NT-$G$ pdf (4); its first derivarite is:

$$\frac{\partial}{\partial x}f_G(x) = \phi\left[\tan\left(\pi\left[G(x) - \frac{1}{2}\right]\right)\right]\sec^2\left(\pi\left[G(x) - \frac{1}{2}\right]\right)\pi \cdot T(x)\,, \tag{7}$$

where

$$T(x) = g^2(x)\tan\left(\pi\left[G(x) - \frac{1}{2}\right]\right)\left(2 - \sec^2\left(\pi\left[G(x) - \frac{1}{2}\right]\right)\right)\pi + g'(x)$$

and $g'(x)$ is the derivative of $g(x)$. Since the terms multiplying $T(x)$ in (7) are positive, we have $\frac{\partial}{\partial x}f_G(x) = 0 \Leftrightarrow T(x) = 0$. In other words, the critical points are the roots of $T(x)$. If the sign of the second derivative of (4) evaluated at a particular critical point is non-positive, then the point is a mode. The explicit expression for $\frac{\partial^2}{\partial x^2}f_G(x)$ is presented in appendix A.

### 2.3. Special submodels from the normal-tangent-$G$ class

Here we discuss two distributions from the NT-$G$ class.

### 2.3.1. Normal-tangent-Weibull distribution

The Weibull distribution is well-known by practitioners of diverse science fields. It is widely used in reliability analysis but also in hydrology and oceanography modelling data. For example, in recent paper Huang and Dong (2019) constructed probability distributions of wave periods in mixed sea states using a mixture Weibull distribution. Given that $G_W(x) = 1 - e^{-(x/\lambda)^k}$ (for $x \geq 0$ and $k, \lambda > 0$) is the Weibull cdf, one can obtain the cdf $F_{NTW}(x|k,\lambda) = \Phi\left[\tan\left(\pi\left[\frac{1}{2} - e^{-(x/\lambda)^k}\right]\right)\right]$ of the NT-Weibull distribution replacing $G$ in equation (3) by $G_W$. Likewise, using (4), one gets to the corresponding pdf, which is given below:

$$f_{NTW}(x|k,\lambda) = \phi\left[\tan\left(\pi\left[\frac{1}{2} - e^{-(x/\lambda)^k}\right]\right)\right] \sec^2\left(\pi\left[\frac{1}{2} - e^{-(x/\lambda)^k}\right]\right) \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}.$$

The Weibull hrf may be constant, increasing or decreasing, depending on the value of the shape parameter $k$. On the other hand, the NT-Weibull hrf presents new non-monotonic shapes, as we can see in Figure 2. Notice yet that there are some values of the parameters, for which the pdf is bimodal (Figure 1).

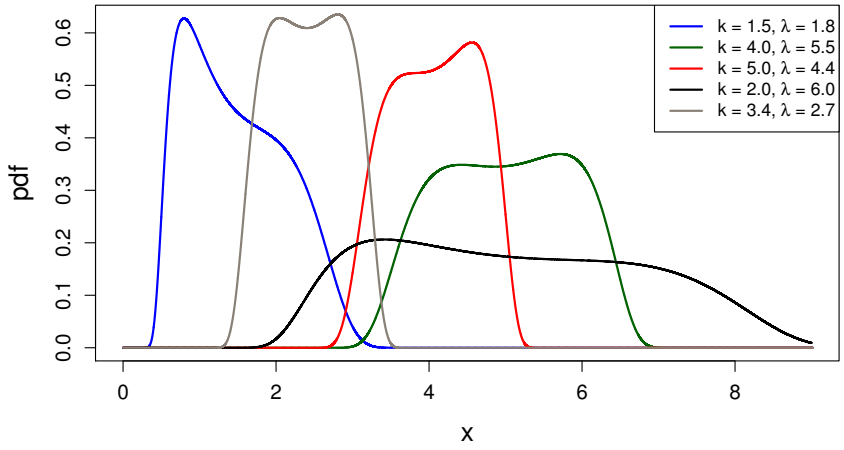


Figure 1: Normal-tangent-Weibull pdf.

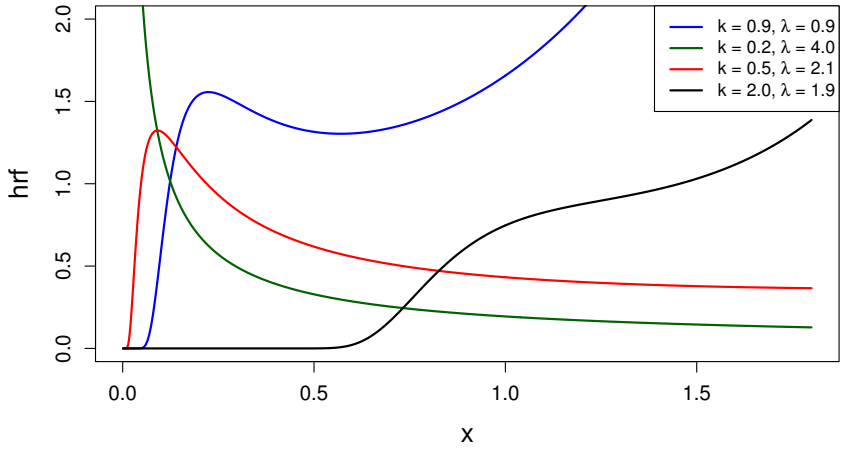


Figure 2: Normal-tangent-Weibull hrf.

Figure 3 shows the graph of the implicit function obtained by the equation $T(x) = 0$ (see section 2.2) for the NT-

Weibull distribution, considering $\lambda = 5.5$; we can see the range of values of $k$ such that the pdf is bimodal. For example, the gray line representing $k = 4$ intersects the graph at the critical points, namely, $x_1 = 4.424524$, $x_2 = 4.860704$ and $x_3 = 5.718673$, where $x_1$ and $x_3$ are the modes and $x_2$ is a local minimum. The green density also illustrates it in Figure 1.

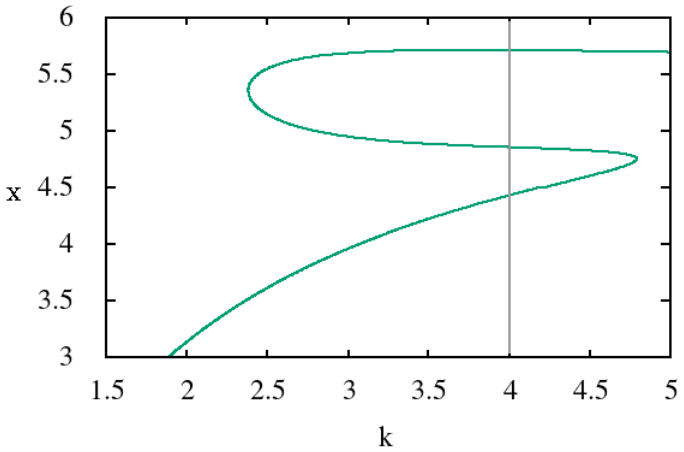

Figure 3: Critical points of the normal-tangent-Weibull pdf for $\lambda = 5.5$.

### 2.3.2. Normal-tangent-log-logistic distribution

The log-logistic distribution is commonly used in survival analysis, but also in different areas like flood frequency analysis (Ahmad et al., 1988), economics and actuarial sciences (Kleiber and Kotz, 2003), where it is known as Fisk distribution. Some important extensions of the log-logistic distribution were used to model cancer data. Among them, one may cite the beta-log-logistic (Lemonte, 2014), used to model the remission times of bladder cancer patients and the McDonald-log-logistic, which was applied by Tahir et al. (2014) to study the survival times of patients with breast cancer. The log-logistic cdf is $G_{LL}(x|\alpha,\beta) = 1-\left(1+(x/\alpha)^{\beta}\right)^{-1}$; replacing $G$ in equation (3) by $G_{LL}$, we get to the NT-log-logistic cdf $F_{NTLL}(x|\alpha,\beta) = \Phi\left[\tan\left(\pi\left[\frac{1}{2}-\left(1+(x/\alpha)^{\beta}\right)^{-1}\right]\right)\right]$, for $x \geq 0$ where $\alpha,\beta > 0$. According to (4), its corresponding pdf is:

$$f_{NTLL}(x|\alpha,\beta) = \phi\left[\tan\left(\pi\left[\frac{1}{2}-\left(1+(x/\alpha)^{\beta}\right)^{-1}\right]\right)\right]\sec^2\left(\pi\left[\frac{1}{2}-\left(1+(x/\alpha)^{\beta}\right)^{-1}\right]\right)\frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1+(x/\alpha)^{\beta})^2}\pi\,.$$

We bring a simulation study and an application of $f_{NTLL}$ to a dataset related to a breast cancer biomarker in sections 3 and 4. Plots of the NT-log-logistic pdf and hrf are exhibited in Figures 4 and 5 respectively, considering different values of $\alpha$ and $\beta$. It is worth remarking that the distribution can have different shapes of hrf, such as inverse "J", unimodal or increasing.

## 2.4. Series representation

The standard normal cdf may be linearly represented by:

$$\Phi(z) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\sum_{n=0}^{\infty}\left(-\frac{1}{2}\right)^n \frac{z^{2n+1}}{n!(2n+1)}\,. \tag{8}$$

Using the result of equation (8) in equation (3), we have:

$$F_G(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\sum_{n=0}^{\infty}\frac{(-1/2)^n}{n!(2n+1)}\left[\tan\left(\pi\left[G(x)-\frac{1}{2}\right]\right)\right]^{2n+1}. \tag{9}$$

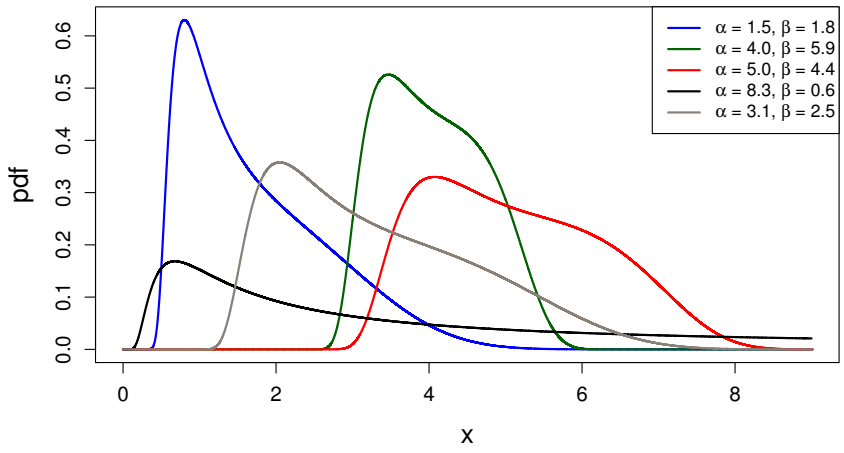


Figure 4: Normal-tangent-log-logistic pdf.

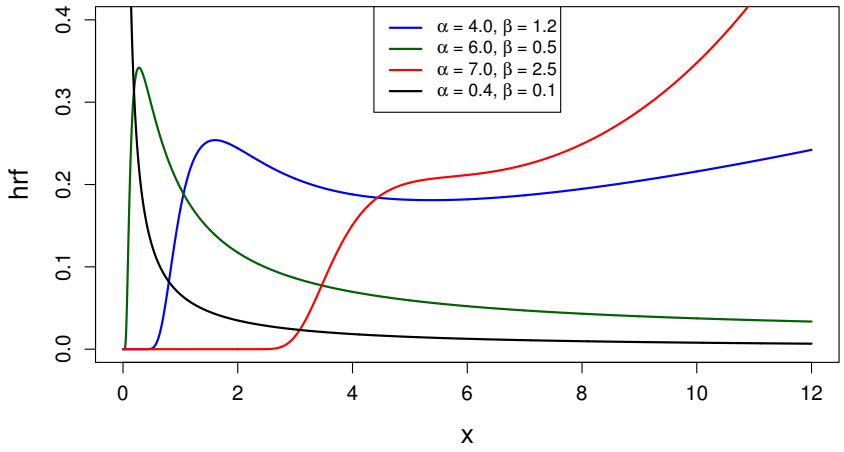


Figure 5: Normal-tangent-log-logistic hrf.

The Maclaurin series of the tangent function is given by:

$$\tan(z) = \sum_{m=1}^{\infty} \frac{(-1)^{m-1} 2^{2m} (2^{2m} - 1) B_{2m}}{(2m)!} z^{2m-1}, \tag{10}$$

for $|z| < \pi/2$, where $B_m$ are the Bernoulli numbers, such that $B_0 = 1$ and $B_m = -\sum_{r=0}^{m-1} \binom{m}{r} \frac{B_r}{m-r+1}$ for $m > 0$. Using the result of equation (10), the expression inside the wider brackets in equation (9) can be written as a series; that is:

$$\tan\left(\pi\left[G(x) - \frac{1}{2}\right]\right) = \sum_{m=1}^{\infty} a_{m-1} \left(G(x) - \frac{1}{2}\right)^{2m-1}, \tag{11}$$

such that $a_{m-1} = \pi^{2m-1}(-1)^{m-1}2^{2m}(2^{2m}-1)B_{2m}/((2m)!)$. Making $m = k+1$ and then inserting (11) in (9), we get to:

$$F_G(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{(-1/2)^n}{n!(2n+1)} \underbrace{\left[\sum_{k=0}^{\infty} a_k \left(G(x) - \frac{1}{2}\right)^{2k+1}\right]^{2n+1}}_{\text{A1}} . \tag{12}$$

A well-known result on power series raised to integer powers states that:

$$\left(\sum_{k=0}^{\infty} a_k y^k\right)^N = \sum_{k=0}^{\infty} c_k y^k \,, \tag{13}$$

where $c_0 = a_0^N$, $c_k = \frac{1}{ka_0}\sum_{s=1}^{k}(sN-k+s)a_s c_{k-s}$ for $k \geq 1$ and $N \in \mathbb{N}$. Setting $N = 2n+1$, $y = \left[G(x) - \frac{1}{2}\right]^2$ and using (13), the expression A1 in equation (12) can be written as follows:

$$\begin{aligned} \text{A1} &= \left[G(x) - \frac{1}{2}\right]^{2n+1} \left[\sum_{k=0}^{\infty} a_k \left(\left[G(x) - \frac{1}{2}\right]^2\right)^k\right]^{2n+1} \\ &= \left[G(x) - \frac{1}{2}\right]^{2n+1} \sum_{k=0}^{\infty} c_k \left[G(x) - \frac{1}{2}\right]^{2k} \\ &= \sum_{k=0}^{\infty} c_k \left[G(x) - \frac{1}{2}\right]^{2(k+n)+1} , \end{aligned} \tag{14}$$

such that $c_0 = a_0^{2n+1}$ and $c_k = \frac{1}{ka_0}\sum_{s=1}^{k}(2s[n+1]-k)a_s c_{k-s}$ for $k \geq 1$. Now using the binomial theorem, we can rewrite (14) as:

$$\text{A1} = \sum_{k=0}^{\infty} \sum_{j=0}^{2(k+n)+1} c_k \binom{2(k+n)+1}{j} \left(-\frac{1}{2}\right)^{2(k+n)+1-j} G(x)^j \,. \tag{15}$$

Replacing A1 in the equation (12) by the expression in (15), we get to:

$$F_G(x) = \frac{1}{2} + \sum_{n,k=0}^{\infty} \sum_{j=0}^{2(k+n)+1} \underbrace{c_k \frac{(-1/2)^{2k+3n+1-j}}{n!(2n+1)\sqrt{2\pi}} \binom{2(k+n)+1}{j}}_{\delta_{n,k,j}} G(x)^j \tag{16}$$

Finally, using Fubini's theorem on differentiation we can write the derivative of (16) with respect to $x$:

$$f_G(x) = \sum_{n,k=0}^{\infty} \sum_{j=0}^{2(k+n)+1} \delta_{n,k,j} g_j(x) \,, \tag{17}$$

where $\delta_{n,k,j} \in \mathbb{R}$ and $g_j(x) = jG(x)^{j-1}g(x)$ is the pdf of a random variable from the exponentiated family (Mudholkar and Srivastava, 1993). Thus, we can say that (17) is the NT-$G$ pdf (4) expressed as a linear combination of pdfs of exponentiated distributions. This result allows that many important quantities that depend on the pdf may also be represented by a series expansion, like the raw moments, the moment generating function, the characteristic function, the Rényi entropy and the order statistics.

### 2.5. Inference

The maximum likelihood method was used to determine the estimates of the parameters of the NT-log-logistic distribution in the next two sections. Its interesting properties such as asymptotic normality, efficiency and consistency

motivate us using it. Here we present the estimates considering the general pdf (4): let $X$ be a random variable following a NT-$G$ distribution, such that $G(x|\boldsymbol{\theta}) = G_{\boldsymbol{\theta}}(x)$ is the baseline, $g(x|\boldsymbol{\theta}) = g_{\boldsymbol{\theta}}(x)$ is its corresponding pdf and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$ is the $m \times 1$ vector of parameters; given that $\mathbf{X} = (x_1, \ldots, x_n)$ is a complete random sample of size $n$ from $X$, the log-likelihood function is:

$$\ell(\boldsymbol{\theta}|\mathbf{X}) = \sum_{k=1}^{n} \log \phi \left[ \tan \left( \pi \left[ G_{\boldsymbol{\theta}}(x_k) - \frac{1}{2} \right] \right) \right] + 2 \sum_{k=1}^{n} \log \sec \left( \pi \left[ G_{\boldsymbol{\theta}}(x_k) - \frac{1}{2} \right] \right) + n \log \pi + \sum_{k=1}^{n} \log g_{\boldsymbol{\theta}}(x_k)$$

and the score vector is $U(\boldsymbol{\theta}|\mathbf{X}) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\mathbf{X}) = (u_i)_{1 \leq i \leq m}$, where:

$$u_i = \pi \sum_{k=1}^{n} \tan \left( \pi \left[ G_{\boldsymbol{\theta}}(x_k) - \frac{1}{2} \right] \right) \sec^2 \left( \pi \left[ G_{\boldsymbol{\theta}}(x_k) - \frac{1}{2} \right] \right) \frac{\partial}{\partial \theta_i} G_{\boldsymbol{\theta}}(x_k)$$
$$+ 2\pi \sum_{k=1}^{n} \tan \left( \pi \left[ G_{\boldsymbol{\theta}}(x_k) - \frac{1}{2} \right] \right) \frac{\partial}{\partial \theta_i} G_{\boldsymbol{\theta}}(x_k) + \sum_{k=1}^{n} \frac{1}{g_{\boldsymbol{\theta}}(x_k)} \frac{\partial}{\partial \theta_i} g_{\boldsymbol{\theta}}(x_k) \, .$$

One can obtain the MLEs by solving the system of equations $u_i = 0$, for $i = 1, \ldots, m$. For interval estimation of $\boldsymbol{\theta}$, one requires the information matrix $J(\boldsymbol{\theta}|\mathbf{X})$ (see appendix B).

Under standard conditions of regularity, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ follows approximately a multivariate normal distribution $N_m(\mathbf{0}_m, \mathcal{I}_{\boldsymbol{\theta}}{}^{-1})$, where $\mathbf{0}_m$ is an $m \times 1$ vector of zeros and $\mathcal{I}_{\boldsymbol{\theta}}$ is the expectation of $J(\boldsymbol{\theta}|\mathbf{X})$, that is, the expected Fisher information matrix.

## 3. Simulation study

The software R version 3.4.4 (R Core Team, 2018) was used to perform the Monte Carlo simulation study. At each of the 10,000 replications, we generated samples of sizes $n = 50, 100, 200, 500$ from the NT-log-logistic distribution using the quantile function and a uniform random number generator as mentioned in section 2.1. We considered five different values for the parametric vector $(\alpha, \beta)$, as shown in second and third columns of Table 1, and calculated the bias and the mean squared error (MSE) of the estimates under the maximum likelihood method as follows:

$$\text{Bias}_\alpha = \frac{1}{10000} \sum_{j=1}^{10000} (\widehat{\alpha}_j - \alpha) \; , \; \text{Bias}_\beta = \frac{1}{10000} \sum_{j=1}^{10000} \left( \widehat{\beta}_j - \beta \right)$$
$$\text{MSE}_\alpha = \frac{1}{10000} \sum_{j=1}^{10000} (\widehat{\alpha}_j - \alpha)^2 \; , \; \text{MSE}_\beta = \frac{1}{10000} \sum_{j=1}^{10000} \left( \widehat{\beta}_j - \beta \right)^2$$

where $\widehat{\alpha}_j$ and $\widehat{\beta}_j$ are the estimates for $\alpha$ and $\beta$ at the $j$-th replication. We used the simulated annealing technique to find the global maximum of the log-likelihood function. This metaheuristic technique is available in R by the `optim` function, for which one has to define previously a vector of initial values, namely $(\alpha_0, \beta_0)$. To do so, we firstly set $(\alpha_0, \beta_0) = (1, 1)$ and run one single replication considering sample size $n = 50$; thereafter, the estimates produced in this step are assigned to $(\alpha_0, \beta_0)$ and used in all of the scenarios.

As we can see in the fourth column of Table 1, the bias for $\widehat{\alpha}$ is small even in the worst scenario (namely $n = 50$ and $\alpha = 4.5$) where the bias is lesser than 3% of the actual value of the scale parameter; similar behavior occurs for the bias of $\widehat{\beta}$, that is, it is small in all rows of the table. It denotes that the estimates are quite close to the real values of the parameters.

The MSE measures the quality of the estimates and being close to zero is a desirable behavior. Furthermore, one expects the MSE to decrease in as much as the sample size increases; we can see along the rows of sixth and seventh columns of Table 1 that it happens for all of the parametric vectors in study. These results indicate that the MLEs for the NT-log-logistic distribution behave appropriately.

**Table 1: Bias and MSE of the estimates under the maximum likelihood method for the NT-log-logistic model.**

| | Actual value | | Bias | | MSE | |
|---|---|---|---|---|---|---|
| $n$ | $\alpha$ | $\beta$ | $\widehat{\alpha}$ | $\widehat{\beta}$ | $\widehat{\alpha}$ | $\widehat{\beta}$ |
| | 1 | 4 | 0.00046376 | 0.09291362 | 0.00068782 | 0.07479668 |
| | 2.5 | 15 | 0.00010388 | 0.35991811 | 0.0003052 | 1.06129824 |
| 50 | 4.5 | 0.5 | 0.10614282 | 0.01221092 | 0.9864407 | 0.00118524 |
| | 17 | 3.5 | 0.00845304 | 0.08401412 | 0.25989369 | 0.05777012 |
| | 23 | 19 | −0.00233362 | 0.45171175 | 0.01610659 | 1.69783706 |
| | 1 | 4 | 0.00033875 | 0.04688769 | 0.00035208 | 0.0343303 |
| | 2.5 | 15 | 0.00014813 | 0.1792126 | 0.00015622 | 0.48430213 |
| 100 | 4.5 | 0.5 | 0.05093953 | 0.00612755 | 0.47911004 | 0.00054026 |
| | 17 | 3.5 | −0.00346277 | 0.0377612 | 0.13307386 | 0.02597981 |
| | 23 | 19 | 0.00076956 | 0.22728986 | 0.00823883 | 0.77715469 |
| | 1 | 4 | 0.00027705 | 0.0242487 | 0.00017798 | 0.01642673 |
| | 2.5 | 15 | −0.00014381 | 0.09630061 | 0.000079 | 0.23217006 |
| 200 | 4.5 | 0.5 | 0.02681204 | 0.0027474 | 0.23683917 | 0.00025478 |
| | 17 | 3.5 | 0.00504611 | 0.01949051 | 0.06725901 | 0.01249458 |
| | 23 | 19 | −0.00027 | 0.1164786 | 0.00416874 | 0.37098647 |
| | 1 | 4 | 0.00007529 | 0.00964605 | 0.00007166 | 0.00637179 |
| | 2.5 | 15 | 0.00003911 | 0.03382906 | 0.00003185 | 0.08939624 |
| 500 | 4.5 | 0.5 | 0.00916854 | 0.00114182 | 0.09375699 | 0.00009938 |
| | 17 | 3.5 | 0.00116053 | 0.00808462 | 0.02705113 | 0.00487081 |
| | 23 | 19 | 0.0004931 | 0.04661039 | 0.00167963 | 0.14383294 |

## 4. Application

Certainly one of the most significant elements that increase the probability of good results in the treatment against breast cancer is the early detection and an essential ally for the precise diagnosis is the use of biomarkers, namely indicators of the presence or severity of the disease. Santillán-Benítez et al. (2013) evaluated the body mass index and the levels of leptin, leptin/adiponectin ratio and carbohydrate antigen (CA) 15-3 as reliable biomarkers for breast cancer. In this section, we present a real data modelling for the concentration (ng/mL) of the leptin hormone in blood samples collected from 116 women at the University Hospital Centre of Coimbra between 2009 and 2013. The dataset was originally studied by Patrício et al. (2018). We fitted the NT-log-logistic model (2.3.2) to the leptin hormone dataset and compared it to the fits of the log-logistic (LL), normal-log-logistic (NLL), exponentiated log-logistic (ExpLL), beta-log-logistic (BLL), Kumaraswamy-log-logistic (KwLL), Gompertz-log-logistic (GoLL) and McDonald-log-logistic (McDLL); the six latter along the lines of Silveira et al. (2019), Mudholkar and Srivastava (1993), Lemonte (2014), Cordeiro and Castro (2011), Alizadeh et al. (2017) and Tahir et al. (2014), respectively. We calculated the MLEs and the corresponding standard errors (SE) using the function `goodness.fit`, which is part of the R package `AdequacyModel`. This function also determines the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the modified statistics of Anderson-Darling (A*) and Cramér-Von Mises (W*) (Chen and Balakrishnan, 1995).

Since the information criteria are related to the amount of information lost by a given model, one would expect that the smaller the values of AIC and BIC, the better the model. In this sense, we can say that NTLL is the best choice among the other distributions in study (see fourth and fifth columns of Table 2). The NTLL also presents the smallest values of A* and W*. As both statistics indicate the difference between the empirical distribution function and the real underlying cdf, we may say that NTLL fits the leptin hormone dataset better than the other options listed in the first column of Table 2.

Figure 6 displays the histogram of the data and the overlapping pdfs of the three fitted models with the smallest values of A*. At first sight, the three curves seem to be plausible approximations, as they are quite close to the histogram. Notice, though, that the blue curve (NTLL) accommodates to the data slightly better than NLL and GoLL, especially for $x > 35$ ng/mL. We have, therefore, good reasons to point the NTLL as the best alternative, since it outperforms

**Table 2: Fitted distributions to the leptin hormone dataset.**

| Distribution | Parameters | MLE (SE) | AIC | BIC | A* | W* |
|---|---|---|---|---|---|---|
| NTLL | $\alpha$ | 20.9386 (1.104906) | 951.1221 | 956.6293 | 0.1896698 | 0.0208069 |
| | $\beta$ | 1.35599 (0.055317) | | | | |
| NLL | $\alpha$ | 20.9160 (1.288521) | 952.7282 | 958.2353 | 0.2638145 | 0.03833129 |
| | $\beta$ | 0.65997 (0.036861) | | | | |
| LL | $\alpha$ | 20.7503 (1.438392) | 965.2305 | 970.7377 | 0.8680731 | 0.1295943 |
| | $\beta$ | 2.37425 (0.179646) | | | | |
| ExpLL | $\alpha$ | 12.6763 (4.863257) | 965.8206 | 974.0814 | 0.7899317 | 0.1142444 |
| | $\beta$ | 1.96082 (0.253925) | | | | |
| | a | 1.99493 (1.014369) | | | | |
| BLL | $\alpha$ | 6.35073 (13.88986) | 962.1373 | 973.1517 | 0.5208001 | 0.07658672 |
| | $\beta$ | 0.52001 (0.274594) | | | | |
| | a | 21.2451 (25.17988) | | | | |
| | b | 11.6772 (13.20232) | | | | |
| KwLL | $\alpha$ | 6.43204 (5.002918) | 963.3929 | 974.4073 | 0.5783344 | 0.08501646 |
| | $\beta$ | 0.94020 (0.276332) | | | | |
| | a | 6.33240 (3.727066) | | | | |
| | b | 3.98015 (2.797636) | | | | |
| GoLL | $\alpha$ | 8.26329 (2.771076) | 958.7918 | 969.8062 | 0.3448491 | 0.04903762 |
| | $\beta$ | 3.97757 (1.482411) | | | | |
| | a | 0.12551 (0.121035) | | | | |
| | b | 0.21671 (0.065575) | | | | |
| McDLL | $\alpha$ | 1.73046 (2.260548) | 963.9579 | 977.7259 | 0.5112334 | 0.07513361 |
| | $\beta$ | 0.39687 (0.361114) | | | | |
| | a | 17.0467 (13.12941) | | | | |
| | b | 18.0783 (37.67819) | | | | |
| | c | 2.27408 (1.896792) | | | | |

the competing models.

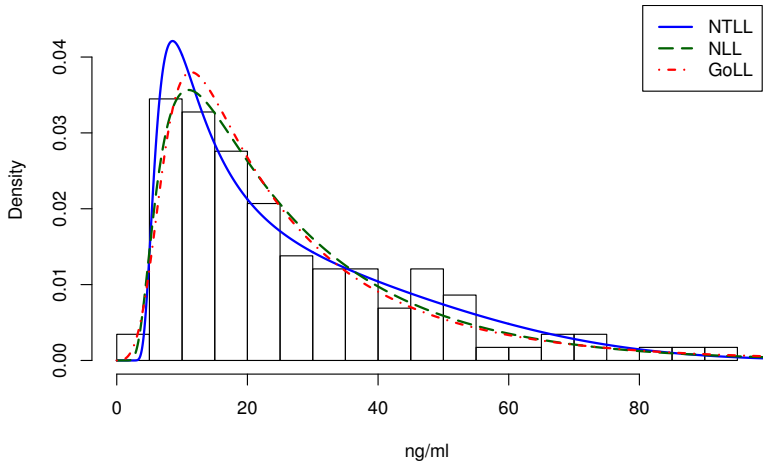


Figure 6: Histogram of leptin concentration dataset and fitted densities.

## 5. Conclusions

A new class of probability distributions called normal-tangent-$G$ is presented, and some of its mathematical properties are discussed, like the quantile function and the series representation of the pdf. The submodels generated by the class are parsimonious; they bring no additional parameters besides those of the baseline $G$ and enjoy the property of identifiability whenever $G$ is identifiable. The class can have unimodal or bimodal distributions with different shapes of pdf and hrf.

Monte Carlo simulation studies are presented to evince the good performance of the MLEs of a NT-$G$ special case, namely, the NT-log-logistic distribution and an application of this particular submodel to a real dataset is carried out to exemplify its modelling benefits. The fitted model is compared to other seven competitive distributions (the baseline itself and six extensions of it) considering AIC, BIC, the Anderson-Darling and the Cramér-von Mises statistics as criteria for goodness-of-fit. All the results attest that the distribution emerged from the NT-$G$ class outperforms the alternative models in study and allows us to point it as a useful and flexible tool for modelling real data.

## References


1. Ahmad, M., Sinclair, C., and Werritty, A. (1988). Log-logistic flood frequency analysis. *Journal of Hydrology*, 98(3):205 – 224.
2. Alizadeh, M., Cordeiro, G. M., Pinho, L. G. B., and Ghosh, I. (2017). The Gompertz-G family of distributions. *Journal of Statistical Theory and Practice*, 11(1):179–207.
3. Brito, C. R., Rego, L. C., Oliveira, W. R., and Gomes-Silva, F. (2019). Method for generating distributions and classes of probability distributions: the univariate case. *Hacettepe Journal of Mathematics and Statistics*, 48(3):897–930.
4. Cakmakyapan, S. and Ozel, G. (2017). The Lindley family of distributions: Properties and applications. *Hacettepe Journal of Mathematics and Statistics*, 46:1113–1137.
5. Chen, G. and Balakrishnan, N. (1995). A general purpose approximate goodness-of-fit test. *Journal of Quality Technology*, 27(2):154–161.
6. Cordeiro, G. M. and Castro, M. (2011). A new family of generalized distibutions. *Journal of Statistical Computation and Simulation*, 81(7):883–898.
7. De Gusmão, F. R. S., Ortega, E. M. M., and Cordeiro, G. M. (2011). The generalized inverse Weibull distribution. *Statistical Papers*, 52(3):591–619.
8. Drapella, A. (1993). The complementary Weibull distribution: unknown or just forgotten? *Quality and Reliability Engineering International*, 9(4):383–385.
9. Huang, W. and Dong, S. (2019). Probability distribution of wave periods in combined sea states with finite mixture models. *Applied Ocean Research*, 92:101938.
10. Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley & Sons, Inc.
11. Lemonte, A. J. (2014). The beta log-logistic distribution. *Brazilian Journal of Probability and Statistics*, 28(3):313–322.
12. Marinho, P. R. D., Cordeiro, G. M., Ramirez, F. P., Alizadeh, M., and Bourguignon, M. (2018). The exponentiated logarithmic generated family of distributions and the evaluation of the confidence intervals by percentile bootstrap. *Brazilian Journal of Probability and Statistics*, 32(2):281–308.
13. Mudholkar, G. S. and Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE transactions on reliability*, 42(2):299–302.
14. Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., and Caramelo, F. (2018). Using resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC cancer*, 18(1):29.
15. R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
16. Santillán-Benítez, J. G., Mendieta-Zerón, H., Gómez-Oliván, L. M., Torres-Juárez, J. J., González-Bañales, J. M., Hernández-Peña, L. V., and Ordóñez-Quiroz, A. (2013). The tetrad BMI, leptin, leptin/adiponectin (L/a) ratio and CA 15-3 are reliable biomarkers of breast cancer. *Journal of clinical laboratory analysis*, 27(1):12–20.
17. Silva, R., Gomes-Silva, F., Ramos, M., Cordeiro, G., Marinho, P. R., and Andrade, T. (2019). The Exponentiated Kumaraswamy-G class: general properties and application. *Revista Colombiana de Estadística*,

42:11–33.

18. Silveira, F. V. J., Gomes-Silva, F., Brito, C. C. R., Cunha-Filho, M., Gusmão, F. R. S., and Xavier-Júnior, S. F. A. (2019). Normal-G class of probability distributions: properties and applications. *Symmetry*, 11:1407.

19. Tahir, M. H., Mansoor, M., Zubair, M., and Hamedani, G. G. (2014). McDonald log-logistic distribution with an application to breast cancer data. *Journal of Statistical Theory and Applications*, 13(1):65–82.

## Appendices

### A. Second derivative of the normal-tangent-$G$ pdf

$$\begin{aligned}\frac{\partial^2}{\partial x^2} f_G(x) = \phi\left[\tan\left(\pi\left[G(x)-\frac{1}{2}\right]\right)\right]\Bigg\{&\pi g''(x)\,\sec^2\left(\pi\left[G(x)-\frac{1}{2}\right]\right)\\ &+\left[-3\pi^2 g(x)\,\sec^4\left(\pi\left[G(x)-\frac{1}{2}\right]\right)+2\pi g(x)(\pi+2)\,\sec^2\left(\pi\left[G(x)-\frac{1}{2}\right]\right)\right]\\ &\times\tan\left(\pi\left[G(x)-\frac{1}{2}\right]\right)g'(x)+\pi^3 g^3(x)\sec^2\left(\pi\left[G(x)-\frac{1}{2}\right]\right)\\ &\times\tan^2\left(\pi\left[G(x)-\frac{1}{2}\right]\right)\left[\sec^4\left(\pi\left[G(x)-\frac{1}{2}\right]\right)-6\,\sec^2\left(\pi\left[G(x)-\frac{1}{2}\right]\right)+4\right]\\ &+\pi^3 g^3(x)\,\sec^4\left(\pi\left[G(x)-\frac{1}{2}\right]\right)\left(2-\sec^2\left(\pi\left[G(x)-\frac{1}{2}\right]\right)\right)\Bigg\}\,,\end{aligned}$$

where $g''(x)$ is the second derivative of $g(x)$.

### B. Normal-tangent-$G$ information matrix

$J(\boldsymbol{\theta}|\mathbf{X}) = -\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}{}^{\top}\ell(\boldsymbol{\theta}|\mathbf{X}) = -(u_{ij})_{1\le i\le m, 1\le j\le m}$, where:

$$\begin{aligned}u_{ij} = \pi^2\sum_{k=1}^{n}&\left[2\,\sec^2\left(\pi\left[G_{\boldsymbol{\theta}}(x_k)-\frac{1}{2}\right]\right)\tan^2\left(\pi\left[G_{\boldsymbol{\theta}}(x_k)-\frac{1}{2}\right]\right)\right.\\ &\left.+\sec^4\left(\pi\left[G_{\boldsymbol{\theta}}(x_k)-\frac{1}{2}\right]\right)\right]\frac{\partial}{\partial\theta_i}G_{\boldsymbol{\theta}}(x_k)\frac{\partial}{\partial\theta_j}G_{\boldsymbol{\theta}}(x_k)\\ &+\pi\sum_{k=1}^{n}\tan\left(\pi\left[G_{\boldsymbol{\theta}}(x_k)-\frac{1}{2}\right]\right)\sec^2\left(\pi\left[G_{\boldsymbol{\theta}}(x_k)-\frac{1}{2}\right]\right)\frac{\partial^2}{\partial\theta_i\partial\theta_j}G_{\boldsymbol{\theta}}(x_k)\\ &+2\pi^2\sum_{k=1}^{n}\sec^2\left(\pi\left[G_{\boldsymbol{\theta}}(x_k)-\frac{1}{2}\right]\right)\frac{\partial}{\partial\theta_i}G_{\boldsymbol{\theta}}(x_k)\frac{\partial}{\partial\theta_j}G_{\boldsymbol{\theta}}(x_k)\\ &+2\pi\sum_{k=1}^{n}\tan\left(\pi\left[G_{\boldsymbol{\theta}}(x_k)-\frac{1}{2}\right]\right)\frac{\partial^2}{\partial\theta_i\partial\theta_j}G_{\boldsymbol{\theta}}(x_k)-\sum_{k=1}^{n}\frac{1}{g^2_{\boldsymbol{\theta}}(x_k)}\frac{\partial}{\partial\theta_i}g_{\boldsymbol{\theta}}(x_k)\frac{\partial}{\partial\theta_j}g_{\boldsymbol{\theta}}(x_k)\\ &+\sum_{k=1}^{n}\frac{1}{g_{\boldsymbol{\theta}}(x_k)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}g_{\boldsymbol{\theta}}(x_k)\,.\end{aligned}$$