Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 2769–2798
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

GS-distributions: A new family of distributions for continuous unimodal variables

J.M. Muiño[a], E.O. Voit[b], A. Sorribas[a,*]

[a]*Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida, Avinguda Rovira Roure, 44, 25198-Lleida, Spain*
[b]*The Wallace H. Coulter Department of Biomedical Engineering at Georgia Tech and Emory University, 313 Ferst Drive, Suite 4103 Atlanta, GA 30322, USA*

Received 21 April 2005; accepted 21 April 2005
Available online 31 May 2005

Abstract

The choice of the best-suited statistical distribution for modeling data is not a trivial issue. Unless a sound theoretical background exists for selecting a particular distribution, one will usually resort to testing various candidates and select a distribution based on its fit to the observed data. While this is a legitimate strategy, it is more objective and efficient to define a sufficiently general family that can be used for this purpose. This approach has a long tradition in statistics, and resulted in various families of distributions, most notably Pearson's. Given such a family, modeling a data set requires estimating the appropriate parameters within this family and assessing the resulting fit. As a contribution to this methodology, the Generalized S-distribution is introduced here as a new family of distributions that can serve as statistical models for unimodal continuous distributions. The article begins with a description of the rationale for defining this family. It then discusses its basic properties and introduces a numerical procedure for determining appropriate parameters using maximum likelihood estimation. Finally, the paper illustrates the distribution and methods with several examples.

Keywords: S-distribution; Recasting; S-system; Distribution family; Maximum likelihood; Estimation; Classification

* Corresponding author. Tel.: +34 973702406; fax: +34 973702426.
E-mail addresses: josem.muinyo@cmb.udl.es (J.M. Muiño), eberhard.voit@bme.gatech.edu (E.O. Voit), albert.sorribas@cmb.udl.es (A. Sorribas).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.04.016

1. Introduction and motivation

The search for families of distributions that can be used as models for empirical data has a long tradition in statistics. The premier example is probably the Pearson family, which contains a number of distributions as special cases (for details see Pearson and Hartley, 1972). Pearson distributions are based on a differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{y(\mu - x)}{a + bx + cx^2}, \tag{1}$$

where y is the density function. For specific parameter values, one obtains various known distributions. For example, $b{=}c{=}0$ and $a > 0$ characterizes the normal distribution. Pearson type I distributions correspond to $b^2/4ac < 0$ and include the beta distribution. Pearson type VII distributions correspond to $b^2/4ac = 0$ and $c > 0$ and include Student's t distribution. Pearson type IV distributions, which correspond to $0 < b^2/4ac < 1$, can represent data distributions with heavy tails and are useful, for instance, for modeling financial and risk management data (Nagahara, 1999). A recent illustrative example of using Pearson's curves for obtaining a distribution for observed data can be found in Podladchikova et al. (2003).

An alternative method for obtaining a model for empirical data was proposed by Johnson (1949). Building upon his ideas, a generalized Johnson family was defined as the set of translation functions comprising any of the forms

$$Z = \gamma + \delta \log\left(\frac{X - \xi}{\xi + \lambda - X}\right), \quad \xi < X < \xi + \lambda, \tag{2}$$

$$Z = \gamma + \delta \log\left(\frac{X - \xi}{\lambda}\right), \tag{3}$$

$$Z = \gamma + \delta \log\left(\frac{X - \xi}{\lambda} + \sqrt{\left(\frac{X - \xi}{\lambda}\right) + 1}\right), \tag{4}$$

where δ, γ, ξ, and λ are parameters, and Z is the standard normal. In practice, the appropriate transformation is obtained by finding the density that matches certain moments or by finding the density that matches certain quantiles (see for instance Hill et al., 1976).

As an alternative to families based on densities, various authors have suggested families based on quantiles (Turner and Pruitt, 1978; Parzen, 1979; Kamps, 1991; Morgenthaler and Tukey, 2000; Jones, 2002, 2004). Yet another alternative was presented by Voit (1992), who proposed a family of distributions known as S-distribution, which has its roots in systems theory (Savageau, 1982). It is defined as

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} = \alpha\left(F(x)^g - F(x)^h\right), \quad F(x_0) = F_0, \tag{5}$$

where $F(x)$ is the cumulative and the parameters satisfy $\alpha > 0$ and $h > g$. The initial condition at x_0 is equivalent to the F_0-quantile. In many applications, F_0 is set equal to 0.5 so that x_0 is the median. Interpreting the left-hand side of Eq. (5) as density, one sees that the density is expressed as a function of the cumulative. This type of functional relationship will be important in the following sections.

Simple rearrangement of (5) leads to the alternative form

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} = \alpha F(x)^g \left(1 - F(x)^k\right), \quad F(x_0) = F_0, \tag{6}$$

with $k > 0$, which is better suited for the generalization proposed in the following. Furthermore, this form will be useful for relating the S-distribution with other closely related representations based on quantiles (see below). For simplicity of discussion, we shall use F to indicate $F(x)$.

As an example, one obtains the S-distribution representation of the exponential as follows. The exponential cumulative is

$$F = 1 - \mathrm{e}^{-\lambda x}. \tag{7}$$

Differentiation of F yields

$$\frac{\mathrm{d}F}{\mathrm{d}x} = \lambda \mathrm{e}^{-\lambda x} = \lambda(1 - F), \tag{8}$$

which exhibits the form of an S-distribution with $g = 0$ and $k = 1$. In a similar manner the logistic distribution

$$F = \frac{1}{1 + \mathrm{e}^{-(x-\alpha)/\beta}} \Rightarrow \frac{1 - F}{F} = \mathrm{e}^{-(x-\alpha)/\beta} \tag{9}$$

can be written in S-distribution form as

$$\frac{\mathrm{d}F}{\mathrm{d}x} = \frac{\mathrm{e}^{-(x-\alpha)/\beta}}{\beta\left(1 + \mathrm{e}^{-(x-\alpha)/\beta}\right)^2} = \frac{1}{\beta} F(1 - F). \tag{10}$$

The generalized exponential distribution defined by Gupta and Kundu (1999) is also included within the S-distribution family. In this case, the cumulative is

$$F = (1 - \mathrm{e}^{-(x-\mu)/\lambda})^\alpha, \tag{11}$$

and

$$\frac{\mathrm{d}F}{\mathrm{d}x} = \frac{\alpha}{\lambda} F^{(\alpha-1)/\alpha} \left(1 - F^{1/\alpha}\right). \tag{12}$$

Except for these cases, the S-distribution does not contain other familiar distributions as special cases, but it has been shown in several analyses that it closely *approximates* most traditional distributions. This facility, which applies to classical continuous as well as discrete distributions often with excellent accuracy (Voit, 1992; Voit and Yu, 1994; Yu and Voit, 1995), identifies the S-distribution as a rather general means for classifying distributions and data, not based on the mathematical formulation of densities but on the basis of shape. Its flexibility, combined with its simplicity of representation, is particularly useful for data modeling and for the development of novel methods of semi-parametric analysis (Balthis et al., 1996; Sorribas et al., 2000, 2002).

The S-distribution has some limitations. First, the number of distributions that are *exactly* represented by S-distributions is limited to the three cases above, and other standard

Table 1
Exact representation of statistical distributions as GS-, S- or Q-distributions

Distribution	GS				S	Q
	α	g	k	γ		
Uniform (a, b)	$1/(b-a)$	0	1	0	−	+
Exponential (λ)	λ	0	1	1	+	+
Generalized Exponential (α, λ, μ)	α/λ	$(\alpha-1)/\alpha$	$1/\alpha$	1	+	−
Logistic (α, β)	$1/\beta$	1	1	1	+	+
$beta(1, b)$	b	0	1	$(b-1)/b$	−	+
$beta(b, 1)$	b	$(b-1)/b$	1	0	−	+
$\mathbf{F}_{2,m}$	1	0	1	$(2+m)/m$	−	+
$\mathbf{F}_{n,2}$	$n^2/4$	$(n-2)/n$	$2/n$	2	−	−

All cases, except for the Generalized Exponential and the $\mathbf{F}_{n,2}$, are included in the Q-distribution, while the S-distribution only includes the exponential and the logistic as special cases.

distributions, notably the normal, are not subsumed in this form. Second, the S-distribution family includes only one symmetric distribution, namely the logistic with $g = 1$ and $k = 1$. Third, the S-distribution family does not readily accommodate distributions with extremely heavy tails, although moderately heavy tails are possible. Finally, S-distributions cannot model finite right tails, as they appear, for instance, in the Beta distribution.

The S-distribution shows formal resemblance with the Q-distribution family, which is defined in terms of quantiles (Turner and Pruitt, 1978; Parzen, 1979; Kamps, 1991; Jones, 2002). This family has appeared in various contexts (see for instance Jones, 2004, and references therein) and can be expressed as:

$$\frac{\mathrm{d}x}{\mathrm{d}F} = \frac{1}{\alpha} F^{-g}(1-F)^{-\gamma}, \quad x\,(F_0) = x_0 \tag{13}$$

with $\alpha > 0$. Inversion of (13) leads to

$$\frac{\mathrm{d}F}{\mathrm{d}x} = \alpha F^{g}(1-F)^{\gamma}, \quad F\,(x_0) = F_0. \tag{14}$$

A number of distributions with analytical cumulative functions can be exactly represented within the Q-distribution family (see Table 1). However, as in the case of S-distributions, there are exceptions such as the generalized exponential, which cannot be represented within the Q-distribution form. Another example is the $\mathbf{F}_{n,2}$ distribution, which is recast as

$$\frac{\mathrm{d}F}{\mathrm{d}x} = \frac{n^2}{4} F^{(n-2)/n}\left(1 - F^{2/n}\right)^2, \quad F(x_0) = F_0. \tag{15}$$

This distribution is close in structure to both the S- and the Q-distributions, but it is not a special case of either one.

The similarity between S- and Q-distributions and the results shown in Table 1 suggest a straightforward extension that subsumes both families. It has the form

$$f(x) = \frac{\mathrm{d}F}{\mathrm{d}x} = \alpha F^{g}\left(1 - F^{k}\right)^{\gamma}, \quad F\,(x_0) = F_0 \tag{16}$$

and ameliorates some of the limitations of the parent distributions, while retaining the advantages of a single, simply structured differential equation. We will demonstrate that this generalization approximates common univariate distributions with even greater accuracy than the Q- and S-distributions. Furthermore, it is better capable of representing symmetric distributions, distributions with finite tails and distributions with very heavy tails. In the following, we give precise definitions for this new family and discuss its basic properties and practical benefits for data modeling and analysis.

2. Generalized S-distribution

2.1. Definition

Based on the motivation in the previous section, we define the generalized S-distribution (GS-distribution) as follows:

Definition 1. The GS-distribution is a continuous, univariate distribution with density

$$f(x) = \frac{\mathrm{d}F}{\mathrm{d}x} = \alpha F^g \left(1 - F^k\right)^\gamma, \quad F(x_0) = F_0. \tag{17}$$

The admissible parameters for the GS-distribution are constrained by the condition

$$\frac{\mathrm{d}F}{\mathrm{d}x} \geqslant 0 \quad \text{for } F \in [0, 1], \tag{18}$$

which necessitates that g and k be positive (real) in order to avoid indetermination at $F = 0$. Similarly, since $\left(1 - F^k\right)^\gamma = 0^\gamma$ at $F = 1$, it follows that $\gamma > 0$. According to these results, $F^g\left(1 - F^k\right)^\gamma \geqslant 0$ for $F \in [0, 1]$, which requires $\alpha > 0$.

If we slightly restrict our considerations to $F \in (0, 1)$, then g and γ may be any real numbers. Notable special cases with $g = 0$ or $\gamma = 0$ arise when we represent some classical distributions as exact cases of GS-distributions (see Table 1 and Section 2.3 for details.) Moreover, negative values of g and γ are allowable, and thus the GS-distribution includes, for instance, Pearson's type I(J) and I(U) distributions as special cases (see Section 2.5).

Remark 1. The GS-distribution can be seen as a location-scale family. Specifically, x_0 is the location parameter corresponding to the F_0-quantile. In many practical cases, it is useful to set $F_0 = 0.5$, which corresponds to specifying the median as x_0. Like the S-distribution, the GS-distribution emphasizes the median, which is determined by the initial value of the differential equation, rather than the mean. Nonetheless, the mean of the GS-distribution can be characterized analytically, as is shown in Appendices B and C.

Parameter α is a scale parameter that plays a similar role as in the S-distribution family (see Voit, 1992; Voit and Schwacke, 1998). It may be written as the product of a scale parameter σ and a normalizing constant c that depends on g, k and γ and renders f a density (see Appendix D for details), i.e.

$$\alpha = c(g, k, \gamma)/\sigma. \tag{19}$$

The parameters g, k, and γ are shape parameters that make this family very flexible in shape. We shall address this point when we discuss the relationship between GS-distribution parameters and moments (see Section 2.5).

Remark 2. For a given x_i, the corresponding value of $F(x_i)$ is obtained by integrating (17) from x_0 to x_i. For this purpose, one may use any integrator, such as the NDSolve procedure in Mathematica. In cases where the right or left tail is finite (see Section 2.6), one must assure that the integration is terminated at the x value that corresponds to the 0th percentile (in the case of a finite left tail) or the 100th percentile (finite right tail).

Remark 3. The GS-distribution has five parameters, namely α, g, k, γ, and the initial condition x_0 for a given F_0), while some of the other families, such as the Pearson family or the Johnson curves discussed in the Introduction, require four or even fewer parameters. In the GS-distribution, the need for five parameters arises to incorporate the exact cases of the S- and Q-distributions, and to expand these families to include other special cases such as the $\mathbf{F}_{n,2}$ distribution, in the form of a $\mathrm{d}F/\mathrm{d}x$ equation. Furthermore, as it is shown in Section 2.3, the shape parameters (g, k, γ) are required to properly approximate a given distribution as a GS-distribution. The x_0 parameter is required as a position parameter, and α is required as a scaling parameter (see Appendix D).

Remark 4. The GS-distribution defined in (17) includes the S-distribution family as the special case $\gamma = 1$, and the Q-distribution family as the special case $k = 1$. Furthermore, it is formally related to the generalized logistic growth function suggested by Tsoularis (2002) and to families of growth functions proposed by Savageau (1980) and Voit (1990).

2.2. Symmetric GS-distributions

In any symmetric distribution

$$P(X \leqslant x_0 - \delta) = P(X \geqslant x_0 + \delta) \quad \forall \delta \geqslant 0, \tag{20}$$

where x_0 is the median, i.e. $F(x_0) = 0.5$. To derive conditions for symmetry in a GS-distribution, it is useful to consider the "F–f plane," where f is plotted against F (Fig. 1). Specifically, the density may be written as $f_F(p) = f\left(F^{-1}(p)\right)$; in other words, the value $f_F(p)$ corresponds to the value of the density at the point $x = F^{-1}(p)$. Defining $P(X \leqslant x_0 - \delta) = p$, we can thus write

$$P(X \leqslant x_0 - \delta) = p = \int_{-\infty}^{x_0-\delta} f(x)\,\mathrm{d}x = \int_0^p f_F(F)\,\mathrm{d}F. \tag{21}$$

On the other hand, since $P(X \leqslant x_0 + \delta) = 1 - p$, we can write

$$P(X \geqslant x_0 + \delta) = p = \int_{x_0+\delta}^{+\infty} f(x)\,\mathrm{d}x = \int_{1-p}^1 f_F(F)\,\mathrm{d}F. \tag{22}$$

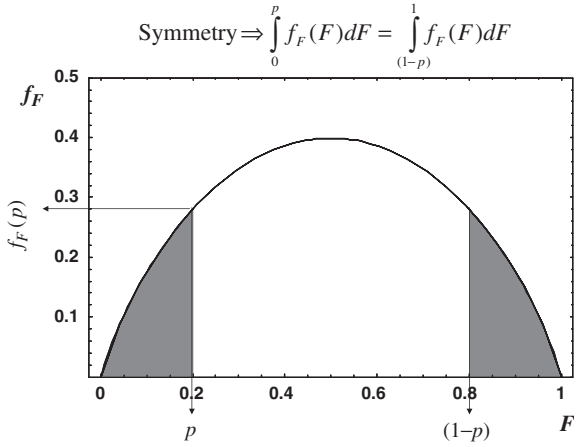

Fig. 1. $F-f$ plane. The GS-distribution pdf corresponding to the logistic distribution is expressed as a function of the cdf. This plane is helpful in discussing the symmetry conditions for any GS-distribution. The shadowed areas represent the integrals indicated in the figure. Symmetry involves the equality of those areas for any value of p.

Thus, symmetry condition (20) may be formulated generically as (see Fig. 1)

$$\int_0^p f_F(F)\,\mathrm{d}F = \int_{1-p}^1 f_F(F)\,\mathrm{d}F \quad \forall p \in [0, 0.5]. \tag{23}$$

In the case of a GS-distribution, this corresponds to

$$\alpha \int_0^p F^g \left(1 - F^k\right)^\gamma \mathrm{d}F = \alpha \int_{1-p}^1 F^g \left(1 - F^k\right)^\gamma \mathrm{d}F \quad \forall p \in [0, 0.5]. \tag{24}$$

Introducing a change of variable, $y = F^k$, the area under f_F between two values p_1 and p_2 is

$$\alpha \int_{p_1}^{p_2} F^g \left(1 - F^k\right)^\gamma \mathrm{d}F = \frac{\alpha}{k} \int_{p_1^k}^{p_2^k} y^{(g+1-k)/k} (1 - y)^\gamma \,\mathrm{d}y. \tag{25}$$

Using this result, symmetry condition in (24) becomes

$$\frac{\alpha}{k} \int_0^{p^k} y^{(g+1-k)/k} (1 - y)^\gamma \,\mathrm{d}y = \frac{\alpha}{k} \int_{(1-p)^k}^1 y^{(g+1-k)/k} (1 - y)^\gamma \,\mathrm{d}y. \tag{26}$$

These integrals correspond to incomplete β functions, which allow us to write the symmetry condition as

$$B_{p^k}\left(\frac{g+1}{k}, 1+\gamma\right) = B\left(\frac{g+1}{k}, 1+\gamma\right) - B_{(1-p)^k}\left(\frac{g+1}{k}, 1+\gamma\right). \tag{27}$$

These results are the motivation for the following theorem.

Theorem 1. *A symmetric GS-distribution is characterized by* $k = 1$ *and* $g = \gamma$.

Proof. The incomplete β function has the following property:

$$B_z(a, b) = B(a, b) - B_{1-z}(b, a). \tag{28}$$

Hence, with $z = p^k$ and taking into account (27), it follows that symmetry in a GS-distribution requires

$$B_{1-p^k}\left(1 + \gamma, \frac{1 + g}{k}\right) = B_{(1-p)^k}\left(\frac{1 + g}{k}, 1 + \gamma\right), \tag{29}$$

which necessitates

$$1 - p^k = (1 - p)^k \quad \forall p \in (0, 0.5) \tag{30}$$

and thus $k = 1$. Furthermore, one can write the generalized incomplete β function as

$$B_{z_1,z_2}(a, b) = B_{1-z_2,1-z_1}(b, a). \tag{31}$$

Hence, if $k = 1$, the following equality must hold:

$$B_{0,p}(g + 1, 1 + \gamma) = B_{1-p,1}(1 + \gamma, g + 1). \tag{32}$$

Symmetry also requires

$$B_{0,p}(g + 1, 1 + \gamma) = B_{1-p,1}(g + 1, 1 + \gamma). \tag{33}$$

Thus, for symmetric distributions, one obtains

$$g + 1 = \gamma + 1 \Rightarrow g = \gamma. \qquad \square \tag{34}$$

Remark. The symmetry condition is easy to understand if one realizes that F and $1 - F$ have the same exponent, so that the approach of the density toward 0 is the same in both tails.

2.3. Approximation of classical distributions as GS-distributions

The GS-distribution contains as exact special cases the distributions indicated in Table 1. In other cases, particular parameter settings in the GS-distribution provide highly accurate approximations of traditional distributions. To assess the quality of the GS-, S-, and Q-distributions approximations, we consider parameters that minimize the Hellinger distance between two densities (ϕ_1, ϕ_2):

$$H\left(\phi_1, \phi_2\right) = \left[\int_{-\infty}^{\infty}\left(\phi_1^{1/2} - \phi_2^{1/2}\right)^2 \mathrm{d}x\right]^{1/2}, \tag{35}$$

Table 2
Parameter values of GS-distribution approximations of classical distributions and corresponding Hellinger distances H

Distribution	α	g	k	γ	H_{GS}	H_S	H_Q
$N(0, 1)$	1.232	0.8379	1.000	0.8379	4.38×10^{-3}	4.30×10^{-2}	4.38×10^{-3}
χ^2_{12}	0.304	0.665	0.762	0.921	6.80×10^{-3}	3.45×10^{-2}	1.34×10^{-2}
χ^2_{23}	0.215	0.722	0.782	0.892	8.10×10^{-3}	4.97×10^{-2}	1.45×10^{-2}
$\chi^2_{12,5}$	0.233	0.689	0.714	0.903	6.58×10^{-3}	6.15×10^{-2}	1.42×10^{-2}
Weibull (20, 1)	19.02	0.943	1.36	0.676	1.16×10^{-3}	6.72×10^{-3}	3.12×10^{-3}
t_5	1.682	1.069	1.000	1.069	7.53×10^{-3}	2.43×10^{-2}	7.53×10^{-3}
t_{15}	1.488	0.893	1.000	0.893	5.47×10^{-3}	2.07×10^{-2}	5.47×10^{-3}
$F_{32,12}$	4.350	0.686	0.554	1.014	1.33×10^{-3}	2.4×10^{-3}	6.14×10^{-3}
$F_{32,32}$	5.421	0.723	0.606	1.004	1.52×10^{-3}	1.79×10^{-3}	5.38×10^{-3}

H_{GS}: GS-distribution approximation. H_S: S-distribution approximation. H_Q: Q-distribution approximation. As the Hellinger distance measures the distance between two densities, a lower value indicates a better approximation. Fig. 2 provides some examples of the fit of classical distributions by GS-distributions. $\chi^2_{12,5}$ is the non-central χ^2 distribution of 12 degrees of freedom and non-centrality parameter 5.

as described in Voit (1992). In this formulation, ϕ_1 represents the target density and ϕ_2 the corresponding GS-, S-, or Q-distribution approximation. The Hellinger distance is chosen here because it is a global measure of agreement between ϕ_1 and ϕ_2. Of course, we could also use other distances, such as the largest difference (in absolute value) between ϕ_1 and ϕ_2.

As Table 2 indicates, the GS-distribution provides excellent representations of traditional distributions with fits that are superior to those obtained with the corresponding S- or Q-distributions. The quality of approximation may be best appreciated in Fig. 2 where we show the agreement between the cumulatives of classical distributions, their approximations as GS-distributions, and the corresponding Q–Q plots. Details of the software used for obtaining the numerical results are provided in Appendix A.

When classical distributions are represented as GS-distributions, limit relationships between distributions are preserved. A good example is provided by distributions corresponding to a GS-distribution with $k = 1$, which includes the important case of symmetric distributions. Fig. 3 illustrates some results in the $g - \gamma$ plane for $k = 1$. First, consider the approach of Student's t toward the normal as the number of degrees of freedom tends to infinity. Each GS-distribution that approximates a specific t distribution corresponds to one point on the line $g = \gamma$, which is situated above the normal. With increasing degrees of freedom these points move closer toward the representation of the normal. Similarly, $\mathbf{F}_{2,m}$ and Beta$(1, b)$ approach the exponential, and the symmetric Beta approaches the point corresponding to the normal. In addition to GS-distributions that correspond to classical distributions, the $g - \gamma$ plane for $k = 1$ contains infinitely many other symmetric distributions. For instance, GS-distributions with $(g, \gamma) > 2$ have infinite mean, and those with $(g, \gamma) > 1.5$ have infinite variance, as we will discuss is Section 2.5. Consequently, the GS-distribution that approximates the Cauchy distribution has $(g = \gamma) > 2$ and $k = 1$. Above these values we would find GS-distribution representations of stable distributions without established names (data not shown).

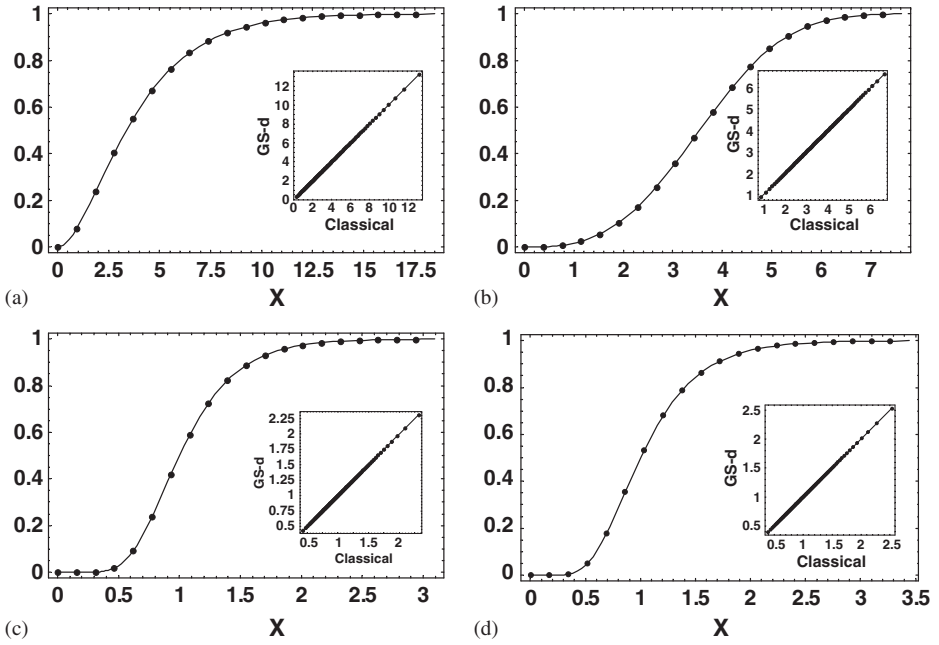

Fig. 2. Approximation of classical distributions with GS-distributions. For each distribution, parameter values of the corresponding GS-distributions were computed by minimization of the Hellinger distance (see text for details). In all cases, the correspondence between the original distribution and its approximation as GS-distribution is excellent. Insets are Q–Q plots. (a) χ^2_4; (b) Weibull (3, 4); (c) $\mathbf{F}_{32,32}$; (d) Lognormal (0, 0.4).

If we restricted our analysis to the $k = 1$ plane, we could conclude that the Q-distribution family would be a sufficient representation. However, while the plane $k = 1$ contains many of the distributions in Table 1, cases like the generalized exponential, the $\mathbf{F}_{n,2}$, and many S-distributions are not included. These distributions require $k = 1/a$, $k = 2/n$, and $k =$ real, respectively, and are thus situated outside the $k = 1$ plane. Furthermore, χ^2 distributions are fitted with higher accuracy if $k < 1$ rather than $k = 1$ (see Table 2). The positions of χ^2 distributions are shown in Fig. 4. Starting with the exponential distribution, the χ^2 distributions with increasing degrees of freedom approach the normal, which as a symmetric distribution is situated in the $k = 1$ plane. These results and the limit relationships between $\mathbf{F}_{n,m}$ distributions and the χ^2_n distribution led to the prediction that $\mathbf{F}_{n,m}$ distributions should also be characterized by $k < 1$. This was indeed confirmed by fitting these distributions with GS-distributions (Table 2). For similar reasons, non-central t distributions are located below the $k < 1$ plane (data not shown.) Finally, other distributions such as the lognormal also require $k < 1$ (data not shown).

Taken together, these results provide justification and support for an extension of S- and Q-distributions to the GS-distribution. This extension requires an extra parameter in either case, but we will show that this additional parameter does not compromise the utility of the resulting family or issues of parameter estimation from observed data.

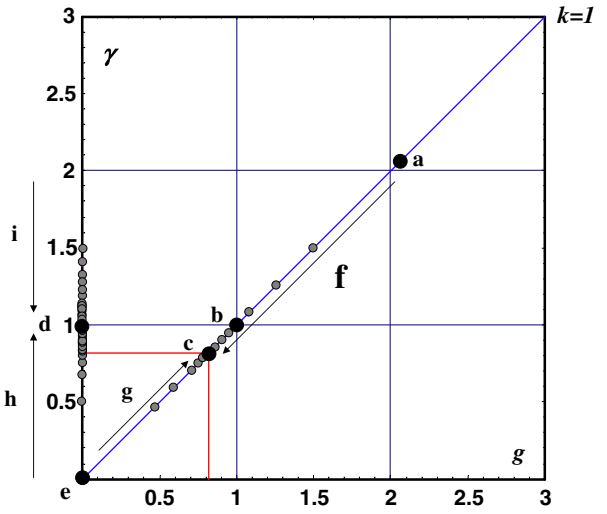

Fig. 3. Some classical distributions represented as GS-distributions with $k = 1$. The relationships between distributions are preserved when represented as GS-distributions: (a) Cauchy distribution; (b) Logistic; (c) Normal; (d) Exponential; (e) Uniform; (f) Starting at point **a** and moving toward point **c**, one locates t distributions with increasing degrees of freedom; (g) Symmetric Beta, starting with Beta, which corresponds to the uniform, and approaching the normal; (h) Beta$(1, b)$, starting with Beta$(1, 1)$, which corresponds to the uniform, and approaching the exponential distribution; and (i) $\mathbf{F}_{2,m}$ with increasing m, starting at $\gamma = 2$ and approaching the exponential.

2.4. Quantiles for GS-distributions

Except for special cases, the GS-distribution does not have an analytical solution for its cumulative. However, it is possible to solve for quantiles, as it was demonstrated for the S-distribution family (Hernández-Bermejo and Sorribas, 2001).

Theorem 2. *Any quantile $F^{-1}(q) = x_q$ of a GS-distribution can be computed as*

$$x_q = x_0 + \frac{B_{F_0^k, q^k}((1-g)/k, 1-\gamma)}{\alpha k}. \tag{36}$$

Proof. It is again useful to subject the GS-distribution in Eq. (17) to a variable transformation of the type $y = F^k$ (Tsoularis, 2002). This yields

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \alpha k y^{(g+k-1)/k}(1-y)^{\gamma}, \tag{37}$$

which is separable:

$$\int_{y_0}^{y_q} (\alpha k)^{-1} y^{(1-g-k)/k}(1-y)^{-\gamma}\,\mathrm{d}y = \left(x_q - x_0\right). \tag{38}$$

The integral on the left-hand side of this equation corresponds to the generalized incomplete Beta function $By_0, y_q((1-g)/k, 1-\gamma)$, divided by (αk), and Eq. (36) follows. □

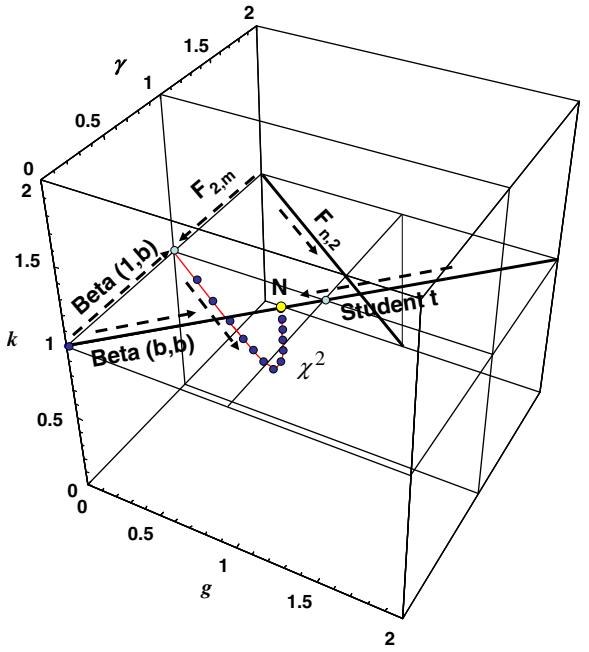

Fig. 4. Localization of statistical distributions in the $g-k-\gamma$ parameter space when represented as GS-distributions. The plane defined by $k=1$ includes all distributions presented in Fig. 3. The $\mathbf{F}_{n,2}$ distribution is located in the $\gamma=2$ plane with $k=2/n$. The χ^2 distribution is approximated by GS-distributions that have $k<1$. With increasing degrees of freedom, the corresponding GS-distributions approach the normal (see text for details).

2.5. Computing moments

Moments of a GS-distribution are obtained as indicated in the following theorem.

Theorem 3. *The jth moment m_j of a GS-distribution may be computed as*

$$m_j = \int_0^1 \left(x_0 + \frac{1}{\alpha k} B_{F_0^k, q^k}\left(\frac{1-g}{k}, 1-\gamma \right) \right)^j \mathrm{d}q. \tag{39}$$

Proof. The jth moment of a random variable is defined as

$$m_j = \int_{-\infty}^{+\infty} x^j f(x,\theta)\,\mathrm{d}x = \int_0^1 x(q)^j\,\mathrm{d}q, \tag{40}$$

where $x(q)$ is the value of the variable that corresponds to $F(x)=q$. When the random variable is GS-distributed, $x(q)$ can be computed from Eq. (36). Direct substitution into (40) leads to (39). □

Some practical results for computing means and for relating the parameter α with the variance are provided in Appendices C and D. As stated in Appendix D (Remark 3), the jth

Table 3
Selected moments computed from GS-distributions that correspond to classical distributions

Moment order	$\mathbf{F}_{2,30}$	$\mathbf{F}_{2,100}$	$\beta(3, 1)$	$\beta(1, 3)$
1	1.07143	1.02041	0.75	0.25
2	2.47253	2.12585	0.60	0.10
3	9.27198	6.78463	0.50	0.05
4	50.5744	29.4984	0.428571	0.0285714
5	379.308	163.88	0.375	0.0178571
6	3793.08	1117.36	0.3333	0.0119048
7	49784.2	9094.82	0.3	0.00833
8	853443	86617.3	0.27272	0.00606
9	1.92925×10^7	950678	0.25	0.004545
10	5.76074×10^8	1.18835×10^7	0.23076	0.003496

All results are exactly the same as those obtained from the original distributions.

moment of a GS-distribution exists and is finite if $(g, \gamma) < (1 + 1/j)$. As a consequence, a GS-distribution has infinite mean if $(g, \gamma) > 2$ and infinite variance if $(g, \gamma) > 1.5$ (see some examples in Figs. 8 and 9.) Table 3 shows examples of moments, computed from Eq. (39), for several distributions that can be represented exactly as GS-distributions and whose parameters fulfill the requirement for finite moments through 10th-order (moments up to m_{10} exist if $(g, \gamma) < 1.10$.) In all cases, Eq. (39) returns precisely the same value as the corresponding moment-generating function for the reference variable.

The results in Fig. 4 show that the (g, k, γ) parameter space of the GS-distribution covers a wide variety of shapes. A complementary point of view on the flexibility of the GS-distribution family may be provided by a characterization of the third and fourth moments of the distributions as suggested by Pearson and Hartley (1972). Fig. 5 exhibits typical curves that relate the GS-distribution parameters to $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$, where μ_j is the central moment of order j. The figure also indicates the regions corresponding to each type of Pearson curve and the GS-distributions covering these regions. Some special cases are worth discussing in more detail. Curve (e) includes the special case of the exponential (point E in the figure) with $\gamma = 1$ and $k = 1$. Starting with the exponential ($g = 0$), curve (e) approaches the logistic (point L, $g = 1$) and continues toward distributions of type IV, for increasing values of g over 1. Curves (a1)–(a3) show the effect of increasing the value of k for a fixed value of γ. A similar effect would be observed for the other cases if we changed the value of k.

Curve (b) includes the special case of the normal distribution (point N). For g values lower than those corresponding to the normal we obtain type I distributions. For values of g higher than those corresponding to the normal we obtain type IV distributions. Finally, cases (c) and (d) correspond to special cases with $\gamma = 0$ and $\gamma = -1$. In (c) the distributions move from type I distributions for negative values of g to the uniform (point U) for $g = 0$ and the exponential for $g = 1$ (point E), respectively. From point U to point E, curve (c) is very close to type IX distributions. Case (d) includes type I(U) distributions for negative values of g. Fig. 5 does not include cases with g or γ over 1.25, which is the limit for computing the 4th-moment in GS-distributions.

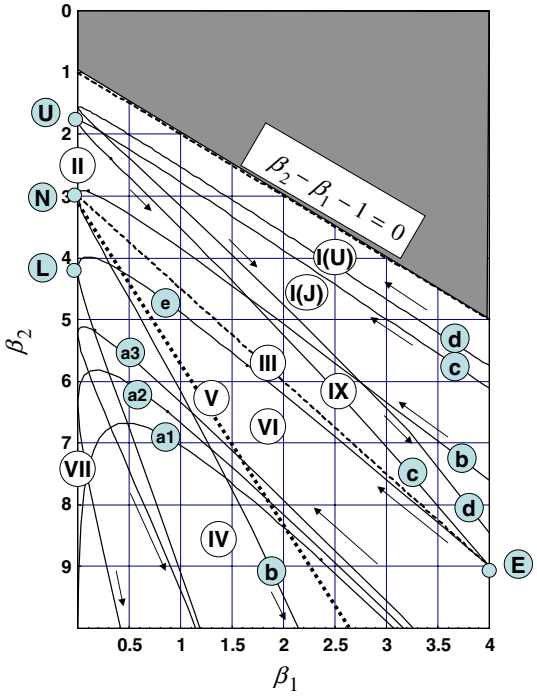

Fig. 5. Moments of GS-distributions. Different lines represent combinations of g, k, γ parameters and their relation to third and fourth moments through: $\beta_1 = \mu_3^2/\mu_2^3$, $\beta_2 = \mu_4/\mu_2^2$, where μ_j is the central moment of order j (see Fig. 7 in Pearson and Hartley, 1972, p. 78). The uppercase letters in circles identify: U: Uniform distribution, N: Normal distribution, L: Logistic distribution, E: Exponential distribution. Roman numbers indicate the different Pearson types. The line $\beta_2 - \beta_1 - 1 = 0$ is the limit for all frequency distributions. Different curves are (see text for details): (a1) $k = 0.5$, $\gamma = 1.1$, g from 0.8 to 1.24; (a2) $k = 1.0$, $\gamma = 1.1$, g from 0.5 to 1.18; (a3) $k = 1.5$, $\gamma = 1.1$, g from 0.3 to 1.14; (b) $k = 1.0$, $\gamma = 0.8379$, g from -0.8 to 1.13; (c) $k = 1.0$, $\gamma = 0.0$, g from -5.5 to 1.00; (d) $k = 3.0$, $\gamma = -1.0$, g from -10.4 to 0.95. The direction of increasing values of g is indicated by an arrow. In each case, the g values are those providing, approximately, values of β_1 and β_2 within the considered range.

Using the first four moments and setting x_0 equal to the median, it is possible to obtain the GS-distribution parameters (α, g, k, γ) by a numerical procedure. First, one computes the mean, variance, skewness, and kurtosis of the data. Then, one searches for the optimal (α, g, k, γ) values that result in computed moments that are as close as possible to the observed mean, variance, skewness, and kurtosis by using an optimization algorithm such as the FindRoot procedure in *Mathematica*. Due to numerical problems this approach does not always converge to a solution. This may be related to the variability of sample moments that compound the numerical procedure. We are currently working on refinements of the numerical methods that deal with this problem. In the samples in which a solution is reached, the final parameter values provide similar fits as those obtained by maximum-likelihood estimation (data not shown.).

2.6. Finite and infinite tails of GS-distributions

The GS-distribution family allows for cases with finite or infinite tails. The following theorem establishes conditions on g and γ that lead to finite tails.

Theorem 4. *A GS-distribution has a finite left tail if $g < 1$. It has a finite right tail if $\gamma < 1$.*

Proof. The quantile equation defined in (36) states

$$x_q = x_0 + \frac{1}{\alpha k} B_{F_0^k, q^k}\left(\frac{1-g}{k}, 1-\gamma\right).$$

The left tail can be characterized as

$$x_{q=0} = x_0 + \frac{1}{\alpha k} B_{F_0^k, 0}\left(\frac{1-g}{k}, 1-\gamma\right). \tag{41}$$

To evaluate this expression, recall the following properties of the incomplete β function:

$$B_{z,0}(a,b) = -B_z(a,b) \quad \text{if } Re(a) > 0, \tag{42}$$

$$B_{z,0}(a,b) = \infty \quad \text{if } Re(a) < 0. \tag{43}$$

Thus, if $g < 1$ the zeroth quantile is given by

$$x_{q=0} = x_0 - \frac{B_{F_0^k}((1-g)/k, 1-\gamma)}{\alpha k} \tag{44}$$

and the GS-distribution has a finite left tail. Otherwise, if $g > 1$, $x_{q=0}$ is infinite. In this case, the left tail is infinite.

For the case of the right tail, we find

$$x_{q=1} = x_0 + \frac{1}{\alpha k} B_{F_0^k, 1}\left(\frac{1-g}{k}, 1-\gamma\right). \tag{45}$$

Considering the property

$$B_{z,1}(a,b) = B(a,b) - B_z(a,b) \quad \text{if } Re(b) > 0, \tag{46}$$

one concludes that a GS-distribution with $\gamma < 1$ has a finite right tail that is given as

$$x_{q=1} = x_0 + \frac{B((1-g)/k, 1-\gamma) - B_{F_0^k}((1-g)/k, 1-\gamma)}{\alpha k}. \tag{47}$$

Furthermore, considering the relationship

$$B_{F_0^k, 1}\left(\frac{1-g}{k}, 1-\gamma\right) = -B_{1-F_0^k, 0}\left(1-\gamma, \frac{1-g}{k}\right), \tag{48}$$

and the properties in Eqs. (42–43), one finds that a GS-distribution has an infinite right tail if $\gamma > 1$. □

Remark 1. The existence of finite tails is of practical relevance for computations involving GS-distributions. For instance, integration below the left endpoint when $g < 1$ leads to computational errors. The same problem appears for the right endpoint when $\gamma < 1$. This is especially important for implementing algorithms for maximum-likelihood estimation, since the parameter search may lead to distributions that have finite tails and thus exclude some data points. The implementation of optimization routines must take this into account.

Remark 2. Distributions with heavy right tails correspond to cases with $\gamma > 1$. As stated in the previous section, when $\gamma > 1.5$ we obtain GS-distributions with infinite variance. Similar results apply for heavy left tails, which are associated with values of g exceeding 1.

3. GS-distributions as models for univariate data

In many practical applications the distribution underlying observed data is unknown (e.g., see Sorribas et al., 2000). As an alternative to naively assuming a particular distribution, the GS-distribution can be employed as a rather general model that closely approximates the unknown distribution. This section illustrates the use of the GS-distribution with two applications. First we discuss parameter estimation from a set of observed data. We will show that the fitted GS-distribution provides a valid parametric model that permits further analyses, such as the computation of moments and quantiles. For a second example, we focus on data with very long tails, which are difficult to represent by any of the common distributions. Very long tails may arise in various practical applications. A typical example is the distribution of resource use or length of stay associated with diseases like HIV/AIDS (e.g. Simpson and Itzler, 1996). Most individuals with the disease require baseline check-ups and medication at an annual cost of a few hundred or a few thousand dollars. However, some individuals contract secondary infections, which require substantial treatment and sometimes extended hospital stays. This type of situation results in a high peak close to zero and an extreme, thin tail reaching into thousands of dollars. Another example was presented by Lu et al. (1998), who tested the effect of tissue plasminogen activator on the recovery from stroke. In this case, the distributions of lesion volume in treatment and control groups, measured by computer tomography, had high peaks close to zero and very long tails, which complicated the analysis because of obvious non-normality. Even various Box–Cox transformations could not normalize these data. We will show that GS-distributions can accommodate these types of long-tailed datasets.

We will also demonstrate that the GS-distribution provides a tool for generating random samples from distributions with unusual shapes or unknown origin. This task must be solved, for instance, in risk assessments based on Monte Carlo simulations, where input variables to a risk scenario are affected by a variety of factors that prevent a theory-based justification of a particular distribution type.

3.1. The GS-distribution as data model: maximum-likelihood estimation

Given a set of observations $x_1, x_2, \ldots, x_n$, the parameters of a GS-distribution can be obtained with a numerical maximum-likelihood procedure. This method is based on ideas

developed for the S-distribution (Schwacke, 2000; Voit, 2000; March et al., 2003) and provides excellent results. Briefly, the method consists of the following steps:

(1) Select a starting set of parameters $(F_0, x_0, \alpha, g, k, \gamma)$.
(2) For each x_i compute the corresponding $F(x_i) = F_i$ value using the parameter set. This F_i value is typically obtained by integrating the GS-distribution differential equation from x_0 until x_i.
(3) Considering that the likelihood in the case of a GS-distribution can be expressed as:

$$L = \prod_{i=1}^{n} f(x_i, \theta) = \prod_{i=1}^{n} \alpha F_i^g \left(1 - F_i^k\right)^{\gamma}, \tag{49}$$

compute the *log-likelihood* function as

$$\log(L) = n \log(\alpha) + g \sum_{i=1}^{n} \log(F_i) + \gamma \sum_{i=1}^{n} \log\left(1 - F_i^k\right). \tag{50}$$

(4) Search for the parameters that maximize the *log-likelihood*.

It the case of a GS-distribution, a numerical procedure must be used for finding the parameter values maximizing $\log(L)$. Given a F_0 (usually 0.5), we obtain the $\left(\widehat{x_0}, \widehat{\alpha}, \widehat{g}, \widehat{k}, \widehat{\gamma}\right)$ values that maximize the *log-likelihood* using an optimization routine. Specifically, we have used the FindMinimum procedure in *Mathematica* after defining the appropriate routine for computing the *log-likelihood*.

As an illustration of the estimation method we present some examples of fitting a GS-distribution to data sets. First, we generated random samples from $N(0, 1)$ and fitted a symmetric GS-distribution using the procedure outlined above. The results are consistent with the normal distribution obtained by estimating μ and σ (Fig. 6). In all cases tested, the GS-distribution yielded a close approximation to the original distribution. Similar results were obtained by computing the first four moments and by applying the method described in the last paragraph of Section 2.5.

As a second example, consider the GS-distributions corresponding to a data set on 24-h urea excretion in elderly males admitted to an intensive care unit (ICU) (Fig. 7). The original data were recorded as part of a larger study for assessing the mortality in the ICU of the University Hospital in Lleida (Spain). In the absence of further information about the underlying distribution, the GS-distribution provides a useful, parametric default model for further data analyses.

Confidence intervals for parameters of the fitted distributions could be obtained, in principle, by computing the Fisher information matrix, whose inverse would provide an estimate of the variance of the estimates. However, in the case of a GS-distribution this matrix must be obtained by approximate numerical methods, which may lead to unstable estimations and thus to inappropriate confidence intervals. As an alternative, we suggest using a bootstrap procedure. Table 4 shows an example of the performance of this method. After generating a random sample of a given distribution, we generated 300 bootstrap samples and fitted a GS-distribution to each sample. Then, we computed the mean and the corresponding confidence intervals using the 300 estimates of each parameter. To appreciate the quality of the fit, we

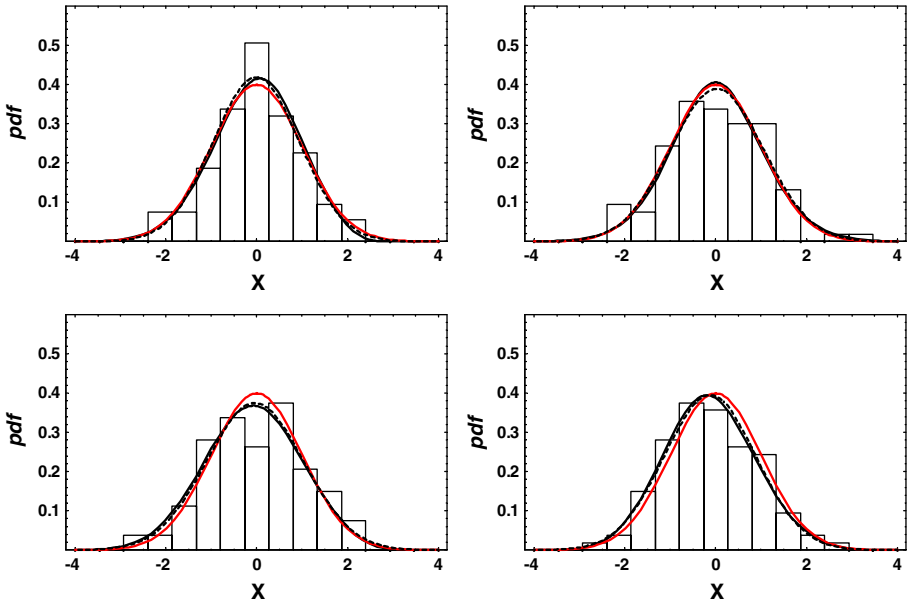

Fig. 6. MLE estimation of GS-distribution parameters. Four samples of size 100 were obtained from N(0,1) (solid gray line), and GS-distributions (solid black line) were fitted by MLE. An alternative fit was obtained by estimating μ and σ under the a priori assumption of normality (dashed line).

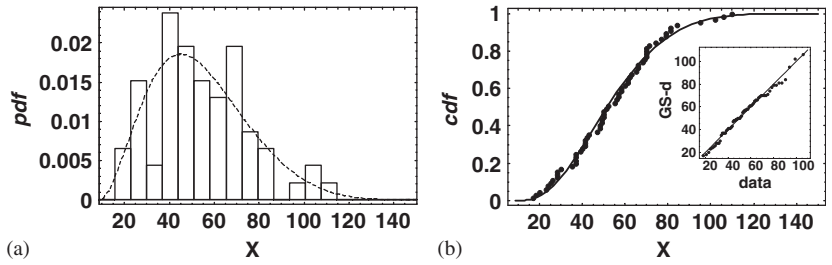

Fig. 7. GS-distribution fitted to data of urea excreted over a period of 24 h in a sample of 78-year-old men ($N = 65$). (a) Frequency histogram and GS-density, (b) Sample cdf and estimated GS-distribution cdf. Inset is a Q–Q plot comparing the fitted GS-distribution and the observed data. The fitted distribution is $\mathrm{GS}\left[F_0, x_0, \alpha, g, k, \gamma\right] = \mathrm{GS}[0.5, 51.49, 0.086, 0.668, 0.403, 0.783]$.

computed the estimation for several quantiles. The results in Table 4 demonstrate that we appropriately recover all original quantiles. Table 4 includes the corresponding estimation of the mean and variance, which are computed, for each bootstrap sample, using the method introduced in Section 2.5.

Table 4
Fitting a GS-distribution to a data set

Parameter	Actual value	Mean of bootstrap samples	95% Confidence interval
x_0	100.0	100.0	(99.6, 100.3)
α	1	0.95	(0.77, 1.12)
g	1	1.06	(0.90, 1.18)
k	1	1.09	(0.64, 1.63)
γ	1	0.94	(0.74, 1.15)
Quantile	**Actual value**	**Mean of bootstrap samples**	**95% Confidence interval**
0.05	97.05	96.60	(95.71, 97.27)
0.10	97.80	97.54	(96.91, 98.09)
0.50	100.00	99.97	(99.63, 100.26)
0.90	102.20	102.01	(101.61, 102.44)
0.95	102.94	102.64	(102.12, 103.13)
Moment	**Actual value**	**Mean of bootstrap samples**	**95% Confidence interval**
Mean	100.0	99.83	(99.48, 100.14)
Variance	3.27	3.64	(2.68, 4.87)

Data were generated by selecting a random sample from a GS[0.5, 100, 1, 1, 1, 1] with $N = 120$. Parameter values were obtained from fitting 300 bootstrap samples. Moments were computed as indicated in Section 2.5 for each bootstrap sample after fitting the GS-distribution parameters.

3.2. Fitting heavy-tail distributions

When data are distributed with heavy tails, it may be difficult to find an appropriate statistical model. In such cases, families of heavy-tailed distributions, such as the Pearson type IV (Nagahara, 1999) or those discussed by Morgenthaler and Tukey (2000), may provide suitable solutions. The GS-distribution also provides a suitable model for this kind of data. For an illustrative set of examples, we first generated data from known GS-distribution with heavy tails. Using the simulated data, we recovered the original distribution with acceptable accuracy, except for some data points at the end of the long tails (Fig. 8). The quality of the obtained fit is assessed in each case by a Q–Q plot.

As an actual example, we used cost data drawn from a large dataset of hospital admissions in 27 US States, describing patients that were admitted for heart failure or heart valve replacement. The data were provided by Dr. Kit N. Simpson of the Medical University of South Carolina. For illustration purposes, we defined two age groups and fitted GS-distributions to the observed data (Fig. 9). The fitted distributions provide an acceptable fit, except at the very end of the long tail. The estimation results show that the fitted parameters are very similar for both groups, which suggests testing if the same distribution could fit both data sets simultaneously. This is accomplished by using the likelihood ratio test after fitting the same distribution to both samples. The fitted parameters for the common distribution

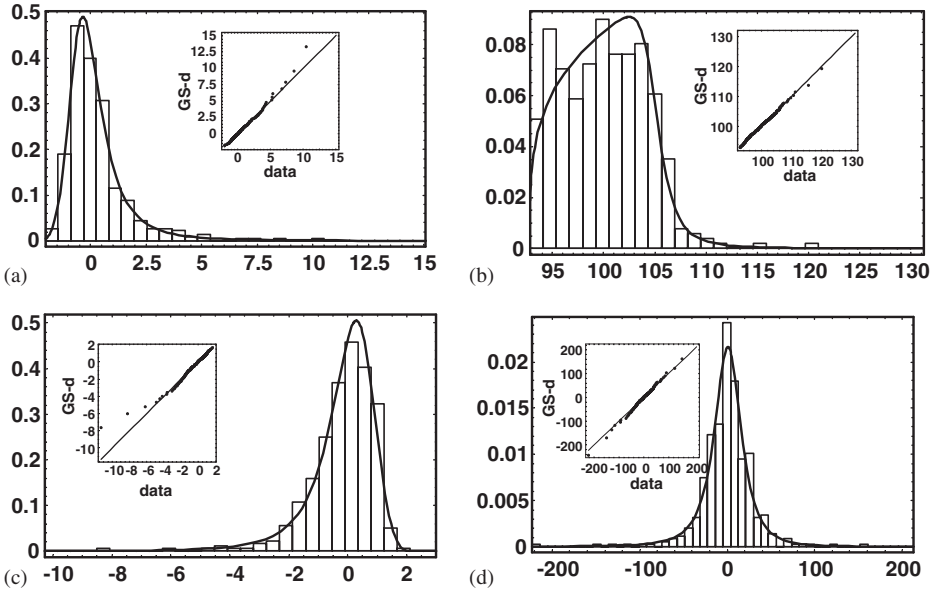

Fig. 8. GS-distribution fitted to simulated data with heavy tails. Q–Q plots (insets) illustrate the quality of fit. Data are simulated from the following distributions GS $[F_0, x_0, \alpha, g, k, \gamma]$: (a) GS [0.5, 0.0, 2.0, 0.8, 1.0, 1.4], (b) GS [0.5, 100.0, 0.1, 0.2, 12.0, 1.9], (c) GS [0.5, 0.0, 2.0, 1.2, 1.0, 0.8], (d) GS [0.5, 0.0, 0.2, 1.5, 1.0, 1.5]. In all cases $n = 400$.

are: $x_0 = 684.97$, $\alpha = 0.0092$, $g = 0.850$, $k = 0.775$, $\gamma = 1.611$. As a result of the test, we can admit a single distribution for both groups ($p = 0.955$).

3.3. Random sample generation using GS-distributions

GS-distributions can be used to generate random samples of univariate unimodal distributions of many shapes. This is accomplished in two steps. First, we obtain a random sample of an uniform distribution on (0, 1) and then transform each random number q in this sample by $x_q = F^{-1}(q)$, which for the GS-distribution corresponds to applying Eq. (36). Our results demonstrate that samples obtained with an approximating GS-distribution are nearly equivalent to samples that were obtained from the original, approximated distribution. Since Eq. (36) applies to all GS-distributions, all we need to do is find suitable GS-distribution parameters for a given distribution and use this equation for generating the required sample. Moreover, one may employ GS-distributions that do not have an analog within the realm of known distributions. For example, one may define a set of q_i values and their corresponding x_q and fit a GS-distribution, which can then be used for generating random samples that approximately agree with the chosen quantiles. Representative examples of random samples from GS-distributions are shown in Fig. 10.

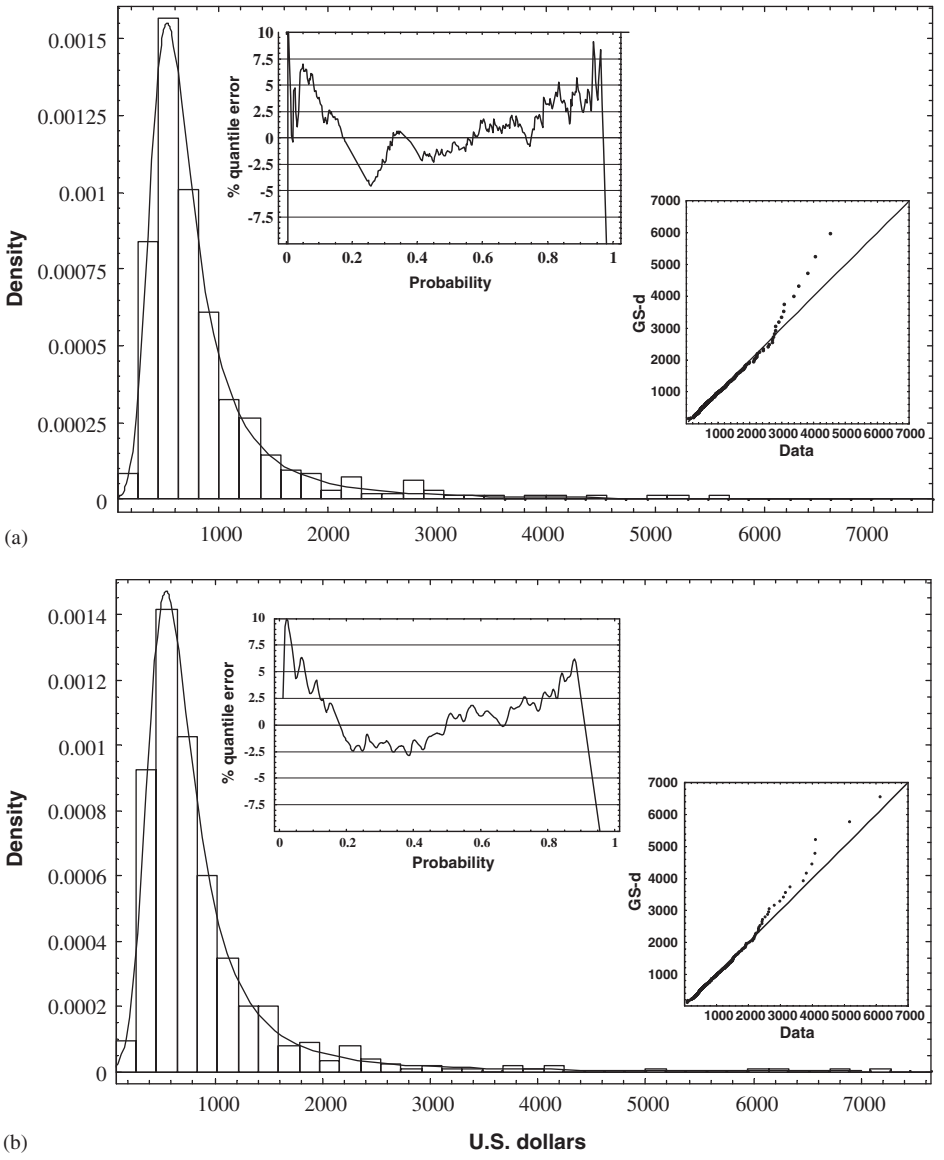

Fig. 9. GS-distributions fitted to cost data associated with patients admitted to a hospital for heart failure or valve replacement. Cost data of this type are characterized by very long tails (see text for details). (a) Age group: 50 to 55 years ($n = 475$). Estimated: GS [0.5, 668.4, 0.010, 0.839, 0.746, 1.634]; (b) Age group: from 55 to 60 years ($n = 642$). Estimated: GS [0.5, 685.0, 0.009, 0.850, 0.775, 1.611]. For each group we include a Q–Q plot and a figure showing the % quantile error between the adjusted GS-distribution and the sample quantile.

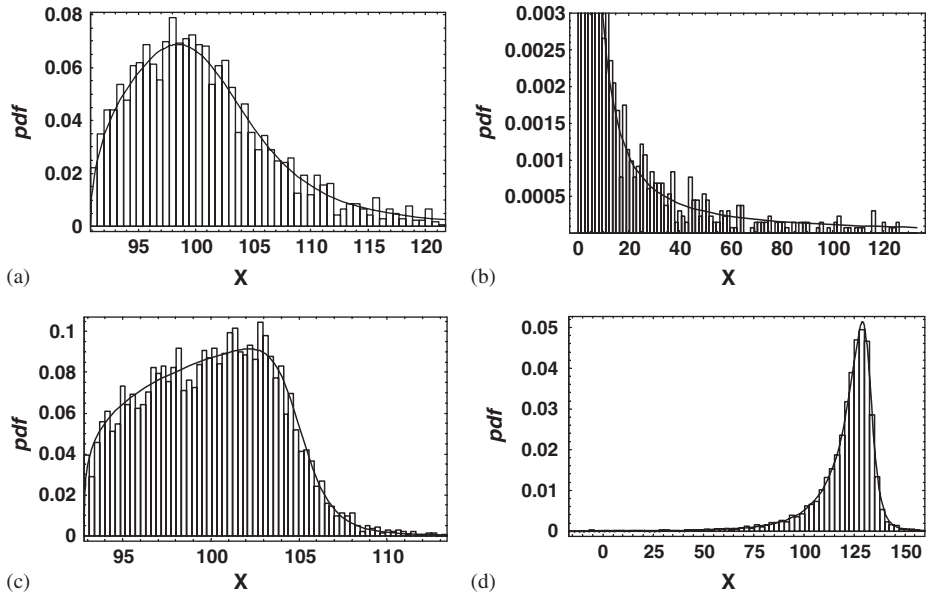

Fig. 10. Random samples of size 5000 obtained from GS-distributions. Parameters for the corresponding GS-distributions, expressed as $\mathrm{GS}\left[F_0, x_0, \alpha, g, k, \gamma\right]$, are: (a) GS [0.5, 100.0, 0.1, 0.3, 3.0, 1.5]; (b) GS [0.001, 0, 100, 1.5, 1.0, 4.0], $n = 10,000$. The vertical axis is truncated to make visible the right tail points. Data shown in the figure span up to the 97% percentile. The maximum value of the sample is 4.28×10^7; (c) GS [0.5, 100, 0.1, 0.2, 12.0, 1.3]; (d) GS [0.5, 125.0, 0.1, 1.2, 5.0, 1.3].

4. Conclusions

The proposed GS-distribution family has several interesting properties. First, it includes as special cases various statistical distributions for which the cumulatives have a closed form. Second, other classical distributions can be closely approximated by GS-distributions. These approximations are more accurate than those based on the S-distribution or the Q-distribution families, which are special cases of the GS-distribution with one less parameter. Third, the GS-distribution can be used to model observed data, when the true underlying distribution is not known.

Our results suggest that the GS-distribution might be a good candidate for a rather general distribution family that can be used for a variety of applications. As a non-standard example, reference intervals can be obtained for an unknown distribution after fitting the appropriate GS-distribution to data. This procedure is an alternative to the use of transformations or non-parametric estimations (cf. Sorribas et al., 2000). The performance of the GS-distributions for these classes of problems remains to be assessed in detail, but the results presented here indicate that the GS-distribution might provide a practical alternative to conventional methods. As another example, hypothesis testing may be based on GS-distributions and a likelihood ratio test. In this case, the resulting methods will provide interesting parametric alternatives to standard non-parametric approaches. Preliminary results on median tests

indicate that the GS-distribution method may be more powerful than the non-parametric alternative. Finally, all methods developed for S-distributions can be easily extended to GS-distributions and are expected to yield more accurate results. Examples here are the estimation of conditional distributions (Sorribas et al., 2000; March et al., 2003) and the evaluation of receiver operating characteristic curves (Sorribas et al., 2002).

Acknowledgements

A. Sorribas and J.M. Muiño are supported by the Fondo de Investigaciones Sanitarias (FIS grant PI020450). The authors are grateful to Dr. Kit N. Simpson for providing data with extreme tails. We also are grateful to two anonymous referees for their suggestions. The final manuscript greatly benefited from their critical reviews.

Appendix A. Computational implementation

All routines required for obtaining the numerical results of this paper have been implemented in a package using the *Mathematica* (Wolfram, 1999) programming language.

We used built-in *Mathematica* functions to compute Beta functions, numerical integration, generate random numbers, etc. Parameter estimation through maximum-likelihood uses the FindMinimum function of *Mathematica*. The GS-distribution was integrated using the NDSolve function for computing the F_i values in the maximum-likelihood estimation routine.

In the future, we plan to implement all procedures in C++. This will speed up some computations and result in a set of reusable routines for those interested in using GS-distributions within other programs. This program will be available soon at www.udl.es/Biomath/GSD.

All figures were also obtained with *Mathematica*.

Appendix B. Computing means in GS-distributions

As a particular consequence of Theorem 3, the mean of a GS-distribution can be obtained as

$$E(X) = m_1 = x_0 + \frac{1}{\alpha k} \int_0^1 B_{F_0^k, q^k}\left(\frac{1-g}{k}, 1-\gamma\right) \mathrm{d}q. \tag{B.1}$$

This equation can be simplified for $g < 2$ and $\gamma < 2$ according to the following theorem.

Theorem B.1. *If $g < 2$ and $\gamma < 2$, the mean of a GS-distribution can be computed as*

$$E(X) = m_1 = x_0 + \frac{1}{\alpha k}\left(B_{F_0^k, 1}\left(\frac{1-g}{k}, 1-\gamma\right) - B\left(\frac{2-g}{k}, 1-\gamma\right)\right). \tag{B.2}$$

Proof. With the variable transformation $u = q^k$, Eq. (B.1) becomes

$$E(X) = x_0 + \frac{1}{\alpha k^2} \int_0^1 u^{1/k-1} B_{F_0^k,u}\left(\frac{1-g}{k}, 1-\gamma\right) \mathrm{d}u. \tag{B.3}$$

Using the following property of the integral of a generalized incomplete Beta function (http://functions.wolfram.com/06.20.21.0008.01)

$$\int z_2^{c-1} B_{z_1,z_2}(a,b)\,\mathrm{d}z_2 = \frac{1}{c}\left(z_2^c B_{z_1,z_2}(a,b) - B_{z_2}(a+c,b)\right). \tag{B.4}$$

Eq. (B.3) becomes

$$E(X) = x_0 + \frac{1}{\alpha k}\left[u^{1/k} B_{F_0^k,u}\left(\frac{1-g}{k}, 1-\gamma\right) - B_u\left(\frac{2-g}{k}, 1-\gamma\right)\right]_0^1. \tag{B.5}$$

This equation leads to indeterminate results for some parameter values. First, when we evaluate this equation at $u = 0$, the result is undefined for $g \geqslant 2$. In this case, $B_0((2-g)/k, 1-\gamma) = \infty$ because $B_0(a,b) = \infty$ if $Re(a) < 0$. Furthermore, using the hypergeometric function ${}_2\mathbf{F}_1(a,b;c;z)$ (Abramowitz and Stegun, 1972; see also the section of the Generalized Hypergeometric Function in http://mathworld.wolfram.com, and references therein), and the property

$$B_{z_1,z_2}(a,b) = \frac{1}{a}\left(z_2^a {}_2\mathbf{F}_1(a, 1-b; a+1; z_2) - z_1^a {}_2\mathbf{F}_1(a, 1-b; a+1; z_1)\right). \tag{B.6}$$

(http://functions.wolfram.com/06.20.26.0004.01) the term $u^{1/k} B_{F_0^k,u}((1-g)/k, 1-\gamma)$ can be expanded for $u = 0$ as

$$\frac{k 0_2^{(2-g)/k} {}_2\mathbf{F}_1((1-g)/k, \gamma; (1-g)/k+1; 0)}{1-g} - \frac{k 0^{1/k} {}_2F_{0_2}^{1-g}\mathbf{F}_1\left((1-g)/k, \gamma; (1-g)/k+1; F_0^k\right)}{1-g}. \tag{B.7}$$

The second term of this expression is equal to 0, while the first term is equal to ∞ if $g \geqslant 2$. Thus, if $g \geqslant 2$, Eq. (B.5) becomes ill-defined: $E(X) = x_0 + \infty - \infty$.

Second, the hypergeometric function ${}_2\mathbf{F}_1(a,b;c;z)$ converges conditionally if

$$-1 < \mathrm{Re}(c-b-a) \leqslant 0,$$

and $z \neq 1$. In (B.7), this condition is

$$-1 < (1-\gamma) < 0 \Rightarrow \gamma < 2.$$

Thus, when $g < 2$ and $\gamma < 2$, Eq. (B.5) reduces to (B.2). □

Remark 1. In some cases, Eq. (B.2) cannot be used because it leads to undefined terms. The situation arises for GS-distributions with $\gamma = 1$, which correspond to S-distributions, or for $\gamma > 1$ and $(1-g)/k = 1$. These particular cases are discussed in Appendix D, along with well-defined expressions for the means.

Remark 2. As a consequence of Theorem B.1, when g or γ are greater than 2, the mean becomes $-\infty$ or $+\infty$.

Remark 3. It can be proved that the jth moment exists and is finite if $(g, \gamma) < (1 + 1/j)$. The same applies for central moments $\mu_j = E((X - E(X))^j)$. Thus, a GS-distribution has an infinite mean if $(g, \gamma) > 2$, and infinite variance if $(g, \gamma) > 1.5$. As expected, in Figs. 3 and 4 the common distributions considered have values of g and γ below these limits. The $\mathbf{F}_{n,2}$ distributions, which have $\gamma = 2$, do not have finite variances. The Cauchy distribution, which has infinite mean and variance, is approximated by a GS-distribution with $(g, \gamma) > 2$. Finally, it is worth indicating that skewness, defined as $\mu_3/\mu_2^{3/2}$ is finite for $(g, \gamma) < 1.33$, while kurtosis, defined as μ_4/μ_2^2 is finite for $(g, \gamma) < 1.25$.

Appendix C. Computation of means of a GS-distribution in special cases

Means of GS-distributions are computed in general as shown in Theorem 3 and Eq. (39). Furthermore, Theorem B.1 provides simplified computations for most cases where $(g, \gamma) < 2$. This part of the Appendix discusses special situations where direct application of Theorem 3 leads to ill-defined quantities.

If $g < 2$ and $\gamma < 2$, Theorem B.1 shows that the mean of a GS-distribution can be computed as

$$m_1 = x_0 + \frac{B_{F_0^k,1}((1-g)/k, 1-\gamma) - B((2-g)/k, 1-\gamma)}{\alpha k}. \tag{C.1}$$

The direct computation of the mean encounters problems in cases where the Beta function becomes infinite. Recalling that

$$\begin{aligned} &B(a, 0) = \infty, \\ &B(0, b) = \infty, \end{aligned} \tag{C.2}$$

and that $\alpha > 0$ and $k > 0$ by definition, it is immediately clear that $B((2-g)/k, 1-\gamma)$ is infinite in the following cases:

(1) $\gamma = 1$,
(2) $(2-g)/k = 0$, which will never occur if $g < 2$.

The case $\gamma = 1$ causes problems since the generalized incomplete $B_{F_0^k,1}((1-g)/k, 0)$ is also equal to ∞, which implies $m_1 = x_0 + \infty - \infty$.

Case I: $\gamma = 1$. This scenario is of particular interest, because it constitutes the special case of an S-distribution. One can resolve issues of indetermination by recalling (B.6) that

$$B_{z,q}(a, b) = \frac{q^a}{a} {}_2\mathbf{F}_1(a, 1-b; a+1; q) - \frac{z^a}{a} {}_2\mathbf{F}_1(a, 1-b; a+1; z). \tag{C.3}$$

Using this relationship transforms the general quantile (36) into

$$x_q = x_0 + \frac{1}{\alpha k} \times \left[\frac{k q^{(1-g)}}{1-g} {}_2\mathbf{F}_1\left(\frac{1-g}{k}, 1; \frac{1-g}{k} + 1; q^k \right) - \frac{k F_0^{(1-g)}}{1-g} {}_2\mathbf{F}_1\left(\frac{1-g}{k}, 1; \frac{1-g}{k} + 1; F_0^k \right) \right]. \tag{C.4}$$

Consequently, substitution of (C.3) into Eq. (B.1)) leads to

$$m_1 = x_0 - \frac{B_{F_0^k}((1-g)/k, 0)}{\alpha k} - \frac{1}{\alpha k} \int_0^1 \frac{u^{(2-g)/k-1}}{1-g} {}_2\mathbf{F}_1\left(\frac{1-g}{k}, 1; \frac{1-g}{k} + 1; u \right) \mathrm{d}u, \tag{C.5}$$

where $u = q^k$. If we consider the property:

$$\int y^{w-1} {}_2\mathbf{F}_1(e, r; t; y)\, \mathrm{d}y = \frac{y^e}{w} {}_3\mathbf{F}_2(e, r, w; t, w+1; y) \tag{C.6}$$

(http://functions.wolfram.com/07.23.21.0003.01) Eq. (C.5) can be written as

$$m_1 = x_0 - \frac{B_{F_0^k}((1-g)/k, 0)}{\alpha k} + \frac{\psi((2-g)/k) - \psi((1-g)/k)}{\alpha k}, \tag{C.7}$$

where ψ is the *digamma* function (Abramowitz and Stegun, 1972).

Case I.a: $\gamma = 1$, $g = 1$. If $\gamma = 1$ and $g = 1$, Eq. (C.7) is ill-defined because: (1) $\psi(-n) = \infty$ if $n \in N$, which occurs when $g = 1$; and (2) $B_{F_0^k}(0, 0) = \infty$ since $B(-n, b) = \infty$ if $n \in N$. In this case, the quantile equation is

$$x_q = x_0 + \frac{B_{F_0^k, q^k}(0, 0)}{\alpha k} = x_0 + \frac{\mathrm{Log}\left(q^k\right) - \mathrm{Log}\left(1 - q^k\right) - \mathrm{Log}\left(F_0^k\right) + \mathrm{Log}\left(1 - F_0^k\right)}{\alpha k}, \tag{C.8}$$

which reduces to

$$m_1 = x_0 + \frac{\gamma_{\mathrm{Euler}} - \mathrm{Log}\left(F_0^k\right) + \mathrm{Log}\left(1 - F_0^k\right) + \psi(1/k)}{\alpha k}. \tag{C.9}$$

In this equation, γ_{Euler} is Euler's constant and has the approximate value 0.57722 (Abramowitz and Stegun, 1972).

Case I.b: $\gamma = 1$, $(1-g)/k = -1$. When $\gamma = 1$ and $(1-g)/k = -1$, the expression (C.7) is ill-defined since $\psi(-n) = \infty$ if $n \in N$ and $B_{F_0^k}(-n, b) = \infty$ if $n \in N$. In this case, the quantile equation becomes

$$\begin{aligned} x_q &= x_0 + \frac{B_{F_0^k, q^k}(-1, 0)}{\alpha k} \\ &= x_0 + \frac{-q^{-k} + F_0^{-k} + \text{Log}\left(q^k\right) - \text{Log}\left(1-q^k\right) - \text{Log}\left(F_0^k\right) + \text{Log}\left(1-F_0^k\right)}{\alpha k}, \end{aligned} \tag{C.10}$$

which reduces to

$$m_1 = x_0 + \frac{\gamma_{\text{Euler}} + F_0^{-k} + (1/(k-1)) - \text{Log}\left(F_0^k\right) + \text{Log}\left(1-F_0^k\right) + \psi(1/k)}{\alpha k}. \tag{C.11}$$

In this equation, a value $k = 1$ would lead to problems. However, this case is not possible since $(1-g)/k = -1$ together with $k = 1$ would imply $g = 2$, and we are only considering GS-distributions with $g < 2$.

Case I.c: $\gamma = 1$, $(1-g)/k = 1$. In this sub-case, the quantile equation reads

$$x_q = x_0 + \frac{B_{F_0^k, q^k}(1, 0)}{\alpha k} = x_0 + \frac{\text{Log}\left(1-F_0^k\right) - \text{Log}\left(1-q^k\right)}{\alpha k}, \tag{C.12}$$

and the mean is

$$m_1 = x_0 + \frac{\psi(1/k+1) + \gamma_{\text{Euler}} + \text{Log}\left(1-F_0^k\right)}{\alpha k}. \tag{C.13}$$

Case II: $\gamma > 1$, $(1-g)/k = 1$. This is a special case that leads to indetermination in (C.1) because of the term

$$B_{F_0^k, 1}(1, 1-\gamma) = \frac{-0^{1-\gamma} + (1-F_0)^{1-\gamma}}{1-\gamma}. \tag{C.14}$$

In this case, the quantile equation can be simplified to

$$x_q = x_0 + \frac{B_{F_0^k, q^k}(1, 1-\gamma)}{\alpha k} = x_0 + \frac{\left(1-F_0^k\right)^{1-\gamma} - \left(1-F^k\right)^{1-\gamma}}{\alpha k(1-\gamma)}, \tag{C.15}$$

and we obtain

$$m_1 = x_0 + \frac{\frac{\left(1-F_0^k\right)^{1-\gamma} k}{1-\gamma} - \frac{\Gamma(1/k)\Gamma(1-\gamma)}{\Gamma(2+1/k-\gamma)}}{\alpha k^2}. \tag{C.16}$$

Symmetric GS-distributions. $k = 1$, $g = \gamma$. This case is well defined but permits a simplified expression for the mean. Specifically, one may write

$$m_1 = x_0 + \frac{B_{F_0, 1}(1-g, 1-g) - B(2-g, 1-g)}{\alpha}, \tag{C.17}$$

which can be rewritten as

$$m_1 = x_0 + \frac{B_{F_0,0.5}(1-g,1-g) + B_{0.5,1}(1-g,1-g) - B(2-g,1-g)}{\alpha}. \tag{C.18}$$

Using the properties of the Beta function:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \tag{C.19}$$

and

$$\Gamma(z+1) = z\Gamma(z), \tag{C.20}$$

one immediately proves that

$$\frac{B(1-g,1-g)}{2} = B(2-g,1-g). \tag{C.21}$$

These results and the property

$$B_{0.5,1}(1-g,1-g) = \frac{B(1-g,1-g)}{2} \tag{C.22}$$

reduce the general expression for the mean of a symmetric GS-distribution to the much simpler formula

$$m_1 = x_0 + \frac{B_{F_0,0.5}(1-g,1-g)}{\alpha}. \tag{C.23}$$

Appendix D. Relationship between parameter α and the variance of a GS-distribution

Results of fitting GS-distributions to well-known classical distributions suggest that parameter α is inversely related to the variance of the variable. For instance, in the case of a normal distribution, we empirically find that α equals $0.282095/\sigma$. Eq. (39) leads to the following result:

Theorem D.1. *The parameter α of a GS-distribution is inversely proportional to the square root of the variance.*

Proof. The variance of a random variable is given as $\mathrm{Var}(X) = m_2 - m_1^2$. According to Eq. (39), the second moment may be computed as

$$m_2 = \int_0^1 \left(x_0 + \frac{1}{\alpha k} B_{F_0^k,q^k}\left(\frac{1-g}{k}, 1-\gamma\right)\right)^2 \mathrm{d}q, \tag{D.1}$$

which can be expressed as

$$m_2 = x_0^2 + \frac{2x_0}{\alpha k} C_1 + \frac{1}{\alpha^2 k^2} C_2^2, \tag{D.2}$$

with

$$C_1 = \int_0^1 B_{F_0^k,q^k}\left(\frac{1-g}{k}, 1-\gamma\right) \mathrm{d}q, \tag{D.3}$$

$$C_2^2 = \int_0^1 \left(B_{F_0^k,q^k}\left(\frac{1-g}{k}, 1-\gamma\right)\right)^2 \mathrm{d}q. \tag{D.4}$$

Furthermore, m_1^2 can be expressed as

$$m_1^2 = x_0^2 + \frac{2x_0}{\alpha k} C_1 + \frac{1}{\alpha^2 k^2} C_1^2. \tag{D.5}$$

These results yield

$$\mathrm{Var}(X) = m_2 - m_1^2 = \frac{\left(C_2^2 - C_1^2\right)}{\alpha^2 k^2} \Rightarrow \alpha = \frac{c(g,k,\gamma)}{\sqrt{\mathrm{Var}(X)}}. \qquad \square \tag{D.6}$$

Remark 1. The proportionality constant $c(g, k, \gamma)$ is a function of the parameters g, k, and γ. From (D.6), $c(g, k, \gamma)$ can be expressed

$$c(g,k,\gamma) = \alpha\sqrt{\left(m_2 - m_1^2\right)}. \tag{D.7}$$

Remark 2. For an alternative proof, which was originally presented for the S-distribution, but applies to the GS-distribution as well, see Voit and Schwacke (1998).

References

Abramowitz, M., Stegun, I.A., 1972. Handbook of Mathematical Functions, 9th Printing. Dover Publications, New York.

Balthis, W.L., Voit, E.O., Meaburn, G.M., 1996. Setting prediction limits for mercury concentrations in fish having high bioaccumulation potential. Environmetrics 7, 429–439.

Gupta, R.D., Kundu, D., 1999. Generalized exponential distributions. Austral. New Zealand J. Statist. 41, 173–188.

Hernández-Bermejo, B., Sorribas, A., 2001. Analytical quantile solution for the S-distribution, random number generation and statistical data modelling. Biometrical J. 43, 1017–1025.

Hill, I.D., Hill, R., Holder, R.L., 1976. Algorithm AS 99: fitting Johnson curves by moments. Appl. Statist. 25, 180–189.

Jones, M.C., 2002. The complementary beta distribution. J. Statist. Plan. Inference 104, 329–337.

Jones, M.C., 2004. Families of distributions arising from distributions of order statistics. Test 13, 1–43.

Johnson, N.L., 1949. Systems of frequency curves generated by methods of translation. Biometrika 36, 149–176.

Kamps, U., 1991. A general recurrence relation for moments of order statistics in a class of probability distributions and characterizations. Metrika 38, 215–225.

Lu, M., Tilley, B.C., Li, S., 1998. Issues on permutation tests: applications in analysis of CT lesion volume in the NINDS T-PA stroke trial. 1998 Proceedings of the Biopharmaceutical Section, American Statistical Association, Alexandria, VA, pp. 27–32.

March, J., Trujillano, J., Tort, M., Sorribas, A., 2003. Estimating conditional distributions using a method based on S-distributions: reference percentile curves for body mass index in Spanish Children. Growth Dev. Aging 67, 59–72.

Morgenthaler, S., Tukey, J.W., 2000. Fitting quantiles: doubling, HR, HQ, and HHH distributions. J. Comp. Graph. Statist. 9, 180–195.

Nagahara, Y., 1999. The PDF and CF of Pearson type IV distributions and the ML estimation of the parameters. Statist. Probab. Lett. 43, 251–264.

Parzen, E., 1979. Nonparametric statistical data modelling (with comments). J. Amer. Statist. Assoc. 74, 105–131.

Pearson, E.S., Hartley, H.O. (Eds.), 1972. Biometrika Tables for Statisticians, vol. 2. Cambridge University Press, Cambridge.

Podladchikova, O., Lefebvre, B., Krasnoselskikh, V., Podladchikov, V., 2003. Classification of probability densities on the basis of Pearson's curves with application to coronal heating simulations. Nonlinear Process. Geophys. 10, 323–333.

Savageau, M.A., 1980. Growth equations: a general equation and a survey of special cases. Math. Biosci. 48, 267–278.

Savageau, M.A., 1982. A suprasystem of probability distributions. Biometrical J. 24, 323–330.

Schwacke, L.H., 2000. SDIST user's manual. Internal Report, Department of Biometry and Epidemiology, Medical University of South Carolina.

Simpson, K.N., Itzler, R., 1996. Exploratory economic analysis of data from randomized clinical trials of antiretroviral therapy in HIV disease populations and recommendations for future study design and analysis protocols. UNC Report to Glaxo Wellcome.

Sorribas, A., March, J., Voit, E.O., 2000. Estimating age-related trends in cross-sectional studies using S-distributions. Statist. Med. 19, 697–713.

Sorribas, A., March, J., Trujillano, J., 2002. A new parametric method based on S-distributions for computing Receiver Operating Characteristic curves for continuous diagnostic tests. Statist. Med. 21, 1215–1235.

Tsoularis, A., 2002. Analysis of logistic growth models. Math. Biosci. 179, 21–55.

Turner, M.E., Pruitt, K.M., 1978. A common basis for survival, growth and autocatalysis. Math. Biosci. 39, 113–123.

Voit, E.O., 1990. S-system analysis of endemic infections. Comput. Math. Appl. 20, 161–173.

Voit, E.O., 1992. The S-distribution: a tool for approximation and classification of univariate, unimodal probability distributions. Biometrical J. 7, 855–878.

Voit, E.O., 2000. A maximum likelihood estimator for shape parameters of S-distributions. Biometrical J. 42, 471–479.

Voit, E.O., Schwacke, L.H., 1998. Scalability properties of the S-distribution. Biometrical J. 40, 665–684.

Voit, E.O., Yu, S., 1994. The S-distribution: approximation of discrete distributions. Biometrical J. 36, 205–219.

Wolfram, S., 1999. The Mathematica Book. fourth ed.. Wolfram Media/Cambridge University Press, Cambridge.

Yu, S., Voit, E.O., 1995. A simple, flexible failure model. Biometrical J. 37, 595–609.