Efficient algorithms for the inversion of the cumulative central beta distribution

A. Gil
Departamento de Matemática Aplicada y CC. de la Computación.
ETSI Caminos. Universidad de Cantabria. 39005-Santander, Spain.

J. Segura
Departamento de Matemáticas, Estadística y Computación,
Univ. de Cantabria, 39005 Santander, Spain.

N.M. Temme
IAA, 1391 VD 18, Abcoude, The Netherlands*

Abstract

Accurate and efficient algorithms for the inversion of the cumulative central beta distribution are described. The algorithms are based on the combination of a fourth-order fixed point method with good non-local convergence properties (the Schwarzian-Newton method), asymptotic inversion methods and sharp bounds in the tails of the distribution function.

1 Introduction

The cumulative central beta distribution (also known as the incomplete beta function) is defined by

$$I_x(p,q) = \frac{1}{B(p,q)} \int_0^x t^{p-1}(1-t)^{q-1}\, dt, \tag{1.1}$$

where we assume that p and q are real positive parameters and $0 \le x \le 1$. $B(p,q)$ is the Beta function

*Former address: Centrum Wiskunde & Informatica (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \tag{1.2}$$

From the integral representation in (1.1) it is easy to check the following relation:

$$I_x(p,q) = 1 - I_{1-x}(q,p). \tag{1.3}$$

In this paper we describe algorithms for solving the equation

$$I_x(p,q) = \alpha,\ 0 < \alpha < 1, \tag{1.4}$$

with p, q given fixed real positive values. In statistical terms, we are computing the quantile function for $I_x(p,q)$.

The beta distribution ia a standard and widely used statistical distribution which has as particular cases other important distributions like the Student's distribution, the F-distribution and the binomial distribution. Therefore, the computational schemes for inverting the central beta distribution can be used to compute percentiles for other distributions related to the beta. For an example for the F-distribution see [1].

The quantile function is useful, for instance, for the generation of random variables following the beta distribution density. In some Monte Carlo simulations the generation of such random variables are required and a massive number of inversions of the beta cumulative distribution are needed. Therefore, it is important to construct methods as reliable and efficient as possible.

Existing algorithms use some simple initial approximations which are improved by iterating with the Newton method. In particular, this is the approach used in the inversion method of the statistical software package **R**, which is based on the algorithm of [9] and the succesive improvements and corrections [4, 2, 3]. In [9], a simple approximation is used in terms of the error function together with two additional starting value approximations for the tails; these initial values are refined by the Newton iteration. As discussed in [4, 2], the Newton method needs some modification to ensure convergence inside the interval $[0,1]$, and further tuning of the Newton method has been considered in recent versions of this algorithm for **R** (but some convergence problem still remain in the present version, as we later discuss).

In this paper the methods for the computation of the inversion of the cumulative beta distribution are improved in several directions. In the first

place, we introduce the Schwarzian-Newton method (SNM) as alternative to Newton's method (NM). With respect to Newton's method the SNM has the advantage of having order of convergence four instead of two. In addition, as explained in [10], the SNM has good non-local properties for this type of functions and it is possible to build an algorithm with certified convergence. In the second place, we analyze initial value approximations (much sharper than those given in [9]) in terms of asymptotic approximations for large values of p and/or q, but which also give accurate values for moderate values; these approximations are given in terms of inverse error functions or the inverse gamma distribution ([12], [6, §10.5.2], [13, §42.3]). We also provide improved approximations for the tails obtained from the sharp bounds described in [11].

An additional direction of improvement of the algorithms is in the selection of the methods of computation of the beta distribution, which are needed in the application of iterative methods (with Newton, SNM or any other choice). This is not discussed in this paper, and we leave this topic for future research. A relatively recent algorithm was given in [5].

2 Methods of computation

We next describe the methods of computation used in the algorithms. First we describe the SNM method, and discuss how a standalone algorithm with certified convergence can be built with this method, provided an accurate method of computation of the beta distribution is available. In the second place we describe the methods for estimating the quantile function based on asymptotics for large p and/or q. Finally, we describe sharp upper and lower bounds for the tails that can be used for solving the problem (1.4) for α close to zero or 1.

2.1 Schwarzian-Newton method

The Schwarzian-Newton method (SNM) is a fourth order fixed point method with good non-local convergence properties for solving nonlinear equations $f(x) = 0$ [10]. The SNM has Halley's method as limit when the Schwarzian derivative of the function $f(x)$ tends to zero.

Given a function $f(x)$ with positive derivative (in our case $f(x) = I_x(p,q) - \alpha$), it is easy to prove that $\Phi(x) = f(x)/\sqrt{f'(x)}$ satisfies the

differential equation

$$\Phi''(x) + \Omega(x)\Phi(x) = 0,\ \Omega = \frac{1}{2}\{f, x\}, \tag{2.1}$$

where $\{f, x\}$ is the Schwarzian derivative of $f(x)$ with respect to x:

$$\{f,\, x\} = \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2. \tag{2.2}$$

The SNM is obtained from the approximate integration of the Riccati equation $h'(x) = 1 + \Omega(x)h(x)^2$, $h(x) = \Phi(x)/\Phi'(x)$ under the assumption that $\Omega(x)$ is approximately constant. In the case of negative Schwarzian derivative (which will be the case for the beta distribution) the iteration function can be written as:

$$g(x) = x - \frac{1}{\sqrt{|\Omega|}}\text{arctanh}\left(\sqrt{|\Omega|}h(x)\right). \tag{2.3}$$

We discuss two implementations: a direct implementation, which gives a convergent algorithm for $p, q > 1$ and an implementation with an exponential change of variables, which is more easy to handle for the rest of cases.

2.1.1 The direct implementation

It is proved in [10] that if $\Omega(x)$ has one and only one extremum at $x_e \in I$ and it is a maximum, then if $\Omega < 0$ the SNM converges monotonically to the root of $f(x)$ in I starting from $x_0 = x_e$. We will use this result for the cumulative central beta distribution, when the parameters p and q are larger than 1. In this case, the function $\Omega(x)$ (the Schwarzian derivative of $f(x)$ with a factor 1/2) is given by

$$\Omega(x) = \frac{(p-1)(q-1)}{2x(1-x)} - \frac{1}{4}\frac{p^2-1}{x^2} - \frac{1}{4}\frac{q^2-1}{(1-x)^2}. \tag{2.4}$$

It is possible to show that for $p > 1$ and $q > 1$, the function $\Omega(x)$ in (2.4) is negative in $(0, 1)$ and has only one extremum (which is a maximum) in that interval. The extremum of $\Omega(x)$ is at

$$x_e = \frac{1}{3\Delta^{1/3}}\frac{(3pq + 3p^2 + 6p)\Delta^{1/3} - \Delta^{2/3} + 3pq\left((p+q)^2 + 8(p+q) + 12\right)}{(p+q)^2 + 2(p+q)}, \tag{2.5}$$

where

$$\Delta = pq\left\{108(p-q)(p+q+1)+27(p^2q+p^3-q^2p-q^3)+3\sqrt{3}(p+q)\right.$$
$$\left.(p+q+2)\sqrt{(p+q+2)(27q+54+q^2p+18pq+27p+p^2q)}\right\}.$$

Then, the fixed point method is (2.3) with $\Omega(x)$ given by (2.4) and

$$h(x)=\frac{f(x)}{\frac{1}{2}\left(-\frac{p-1}{x}+\frac{q-1}{1-x}\right)f(x)+\frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}},$$

where $f(x)=I_x(p,q)-\alpha$.

2.1.2 The exponential implementation

When p and/or q are smaller than 1, it is possible to make a change of variables in order to obtain a negative Schwarzian derivative and simpler monotonicity properties. In particular, with the change of variables

$$z(x)=\log\left(\frac{x}{1-x}\right), \tag{2.6}$$

we obtain that $\Phi(z)=f(z)/\sqrt{\dot{f}(z)}$ (where the dot represents the derivative with respect to z) satisfies $\ddot{\Phi}(z)+\Omega(z)\Phi(z)=0$, where

$$\Omega(z)=\frac{1}{4}\left(-(p+q)(p+q-2)x^2(z)+2(p+q)(p-1)x(z)-p^2\right). \tag{2.7}$$

The function $\Omega(z)$ has its extremum at $z_e=\log(x_e/(1-x_e))$, $x_e=(p-1)/(p+q-2)$. When p and/or q smaller than 1, $\Omega(z)$ can be either be monotonic or it can have a minimum. Convergence of the SNM can be guaranteed, in this case, using the following results [10]: a) if $\Omega(z)$ is negative and decreasing in the interval $I=[a,\,\alpha]$, then the SNM converges monotonically to α for any starting value $z_0\in[a,\,\alpha]$; b) if $\Omega(z)$ is negative and increasing in the interval $I=[\alpha,\,b]$, then the SNM converges monotonically to α for any starting value $z_0\in[\alpha,\,b]$.

The case $p=q=1$ is of course trivial. Apart from this, there are three different cases to be considered:

a) $p\leq 1$, $q>1$: the function $\Omega(z)$ is decreasing. In this case, the SNM uses as starting point a large negative value (in the z variable).

b) $p > 1$, $q \leq 1$: the function $\Omega(z)$ is increasing. In this case, the SNM uses as starting point a large positive value (in the z variable).

c) $p < 1$, $q < 1$: the extremum of $\Omega(z)$ at z_e is reached and it is a minimum. In this case, we use the sign of the function $h(z)$ at z_e to select a subinterval for application of the SNM, according to the previous results. The function $h(z)$ is given by

$$h(z) = \frac{f(x(z))}{-\frac{1}{2}\frac{p-qe^z}{1+e^z}f(x(z)) + \frac{e^{zp}}{B(p,q)(1+e^z)^{p+q}}}. \tag{2.8}$$

When this sign is negative, the SNM uses a large positive value (in the z variable), otherwise the SNM uses a large negative value.

Once the SNM is applied to find the root z_r in the z-variable, the corresponding x-value will be given by $x_r = \frac{e^{z_r}}{1+e^{z_r}}$.

2.1.3 Discussion

We have constructed two methods of order four which are proven to converge with certainty for the initial values prescribed. The method has, in addition, good non-local properties, which means that few iterations are needed for a good estimation of the inverse (typically from 2 to 4 for 20 digits), even without accurate starting values. The exceptions are the tails (α very close to 0 or 1), but we will discuss later how to deal with these cases.

Because the convergence is guaranteed, no special precaution is needed to ensure that the interval $[0, 1]$ in the original variable is not abandoned, as happened with earlier versions of the algorithm given in [9] (see [4]) and as it is still happens for some values in the latests **R** version of this algorithm. For instance, the **R** command qbeta(alpha,600,1.1) does not converge properly if alpha$\in (6.9\,10^{-35}, 1.4\,10^{-20})$. Our method avoids this type of problems.

The performance of the method can be improved by considering initial approximations, which we are discussing next.

2.2 Asymptotic inversion methods

The algorithm considered in [9], which is the basis of the **R** implementation, uses an approximation in terms of the inverse error function, which works

reasonably away from the tails. However, this simple approximation does not give more than two accurate digits, except by accident.

Much more accurate initial approximations (some of them also in terms of error functions) can be obtained from asymptotics for large p and/or q. These are accurate approximations for large and not so large p and/or q, as we later discuss.

This section is based on the results given in [12] and [13, §42.3].

2.2.1 Inversion using the error function

We start with the following representation

$$I_x(p,q) = \tfrac{1}{2}\mathrm{erfc}\left(-\eta\sqrt{r/2}\right) - R_r(\eta), \tag{2.9}$$

where we write $p = r\sin^2\theta$, $q = r\cos^2\theta$ with $0 < \theta < \pi/2$ and η is given by

$$-\tfrac{1}{2}\eta^2 = \sin^2\theta\log\frac{x}{\sin^2\theta} + \cos^2\theta\log\frac{1-x}{\cos^2\theta}. \tag{2.10}$$

When we take the square root for η, we choose $\mathrm{sign}(\eta) = \mathrm{sign}(x-\sin^2\theta)$, this means $\mathrm{sign}(\eta) = \mathrm{sign}(x - p/(p+q))$. In this way, the relation between $x \in (0,1)$ and $\eta \in (-\infty,\infty)$ becomes one-to-one.

Using this representation of $I_x(p,q)$, we solve the equation in (1.4) first in terms of η. When $r = p+q$ is a large parameter, the asymptotic method will provide an approximation to the requested value of η in the form

$$\eta \sim \eta_0 + \frac{\eta_1}{r} + \frac{\eta_2}{r^2} + \frac{\eta_3}{r^3} + \dots . \tag{2.11}$$

The algorithm for computing the coefficients η_i, $i = 0,1,2,\dots$ can be summarized as follows

1. The value η_0 is obtained from the equation

$$\tfrac{1}{2}\mathrm{erfc}\left(-\eta_0\sqrt{r/2}\right) = \alpha. \tag{2.12}$$

2. With $\eta = \eta_0$, equation (2.10) is inverted to obtain a first approximation to the value of x. For inverting this equation, it seems convenient to write it in the form

$$x^p(1-x)^q = \left(\frac{p}{r}\right)^p\left(\frac{q}{r}\right)^q e^{-r\eta^2/2}. \tag{2.13}$$

3. With these values of η_0 and x, the coefficient η_1 is given by

$$\eta_1 = \frac{\log(f(\eta_0))}{\eta_0}, \tag{2.14}$$

where $f(\eta) = \dfrac{\eta \sin\theta \cos\theta}{(x - \sin^2\theta)}$.

4. Higher-order coefficients η_j, $j = 2, 3, \ldots$ can be obtained in terms of x, η_0, η_1, $\sin\theta$ and $\cos\theta$. As an example, the coefficient η_2 is given by

$$\begin{aligned}\eta_2 = &\frac{1}{12\eta_0^3 c^2 s^2 (s^2 - x)^2}\left(s^6\eta_0^2 - \eta_0^2 x^2 - s^4\eta_0^2 - \eta_0^2 s^8 + \right.\\ &12s^6c^2 - 12s^2c^2\eta_1\eta_0^3 x + 12s^2c^2\eta_1\eta_0^3 x^2 - 6\eta_0^2 s^2 c^2 \eta_1^2 x^2 + \\ &12\eta_0^2 s^4 c^2 \eta_1^2 x + 2\eta_0^2 x s^2 + 2\eta_0^2 x s^6 - 6\eta_0^2 s^6 c^2 \eta_1^2 + \\ &12s^2c^2\eta_0^2 x^2 - 12s^2c^2\eta_0^2 x - 2\eta_0^2 x s^4 - \eta_0^2 x^2 s^4 + \\ &\left.\eta_0^2 x^2 s^2 - 24 s^4 c^2 x + 12 s^2 c^2 x^2\right),\end{aligned} \tag{2.15}$$

where $s = \sin\theta$, $c = \cos\theta$.

5. With these coefficients in the expansion (2.11), a value for η is obtained. Then, the inversion of (2.10) will provide the final asymptotic estimation of the x-value.

Using (2.10) we can derive the following expansion or small values of $|\eta|$:

$$\begin{aligned}x = &s^2 + sc\,\eta + \frac{1 - 2s^2}{3}\eta^2 + \frac{13s^4 - 13s^2 + 1}{36sc}\eta^3 + \\ &\frac{46s^6 - 69s^4 + 21s^2 + 1}{270 s^2 c^2}\eta^4 + \ldots,\end{aligned} \tag{2.16}$$

where $s = \sin\theta$, $c = \cos\theta$. For larger values of $|\eta|$, with $\eta < 0$, we rewrite (2.10) in the form $x(1 - x)^\mu = u$, where

$$\mu = \cot^2\theta, \quad u = \exp\left[(-\tfrac{1}{2}\eta^2 + s^2 \ln s^2 + c^2 \ln c^2)/s^2\right], \tag{2.17}$$

and for small values of u we expand

$$\begin{aligned}x = &u + \mu u^2 + \frac{3\mu(3\mu + 1)}{3!}u^3 + \frac{4\mu(4\mu + 1)(4\mu + 2)}{4!}u^4 + \\ &\frac{5\mu(5\mu + 1)(5\mu + 2)(5\mu + 3)}{5!}u^5 + \ldots.\end{aligned} \tag{2.18}$$

A similar approach is possible for positive values of η, giving an expansion for x near unity. In that case we have the equation $x^\nu(1-x)=v$, where

$$\nu=\tan^2\theta,\quad v=\exp\left[(-\tfrac{1}{2}\eta^2+s^2\ln s^2+c^2\ln c^2)/c^2\right], \tag{2.19}$$

and we have the expansion

$$\begin{aligned}1-x=v+\nu v^2+\frac{3\nu(3\nu+1)}{3!}v^3+\frac{4\nu(4\nu+1)(4\nu+2)}{4!}v^4+\\ \frac{5\nu(5\nu+1)(5\nu+2)(5\nu+3)}{5!}v^5+\ldots,\end{aligned} \tag{2.20}$$

The approximations to x obtained in this way will be used for starting the SNM for obtaining more accurate values of x.

2.2.2 Inversion using the incomplete gamma function

In this case, we start from

$$I_x(p,q)=Q(q,\eta p)+R_{p,q}(\eta), \tag{2.21}$$

where $Q(a,x)$ is the incomplete gamma function ratio.

The parameter η is given by

$$\eta-\mu\log\eta+(1+\mu)\log(1+\mu)-\mu=-\log x-\mu\log(1-x), \tag{2.22}$$

where $\mu=q/p$ and x have the following corresponding points:

$$x=0\Longleftrightarrow\eta=+\infty\quad x=1/(1+\mu)\Longleftrightarrow\eta=\mu,\quad x=1\Longleftrightarrow\eta=0. \tag{2.23}$$

So, for $x\in(0,1)$ we have $\operatorname{sign}(\eta-\mu)=-\operatorname{sign}(x-1/(1+\mu))$.

For the representation in (2.21) we assume that p is the large parameter, and we will obtain approximations to the value of η in the form

$$\eta\sim\eta_0+\frac{\eta_1}{p}+\frac{\eta_2}{p^2}+\frac{\eta_3}{p^3}+\ldots,\ p\to\infty. \tag{2.24}$$

We follow similar ideas as in §2.2.1. The value of η_0 can be obtained by solving

$$Q(q,\eta_0 p)=\alpha. \tag{2.25}$$

The inversion of $Q(a,x)$ can be done by using our inversion algorithm described in [7].

Then, a value x_0 is obtained by solving (2.22) for x. With x_0 and η_0 we compute

$$\eta_1 = \frac{\log \phi(\eta_0)}{1 - \mu/\eta_0}, \tag{2.26}$$

where $\phi(\eta)$ is given by $\phi(\eta) = \dfrac{\eta - \mu}{1 - x(1+\mu)} \dfrac{1}{\sqrt{1+\mu}}$.

Other coefficients η_j, $j = 2, 3, \ldots$ can be obtained in terms of μ, x_0, η_0 and η_1.

To compute x from equation (2.22) we can use the inversion algorithm for computing x when in (2.10) η is given. This follows from $\mu = q/p = \cot^2\theta$ and from writing (2.22) in the form

$$\sin^2\theta \left(\mu - \eta + \mu \log \frac{\eta}{\mu}\right) = \sin^2\theta \log \frac{x}{\sin^2\theta} + \cos^2\theta \log \frac{1-x}{\cos^2\theta}. \tag{2.27}$$

This equation can also be written as

$$x^p (1-x)^q = \left(\frac{p}{r}\right)^r e^{q(1+\log\eta) - p\eta}. \tag{2.28}$$

2.3 Interval estimation for the tails

Sharp lower and upper bounds for the solution x of the equation (1.4) in the lower ($\alpha \to 0$) and upper ($\alpha \to 1$) tails of the distribution function can be obtained by using fixed point iterations $x_l^{n+1} = g_l(x_l^n)$ and $x_u^{n+1} = g_u(x_u^n)$, respectively, where the iteration functions $g_l(x)$ and $g_u(x)$ for the lower tail are given by [11]

$$g_l(x) = \left(\alpha B(p,q)\,(p - (p+q)x)\,(1-x)^{-q}\right)^{1/p}, \tag{2.29}$$

and

$$g_u(x) = \left(\frac{\alpha p B(p,q)}{\left(1 + \dfrac{(p+q)}{(p+1)}x + \dfrac{(p+q)(p+q+1)}{(p+1)(p+2)}x^2\right)(1-x)^q} \right)^{1/p}. \tag{2.30}$$

The starting value of the fixed point iterations is $x = 0$. The solution x of the equation (1.4) satisfies $x_l < x < x_u$. These bounds of x can

be used as starting values for the SNM. Notice that, because a lower and an upper bound is obtained we have an estimation of the error for these approximations that can be used to decide if they are accurate enough.

We notice that the approximation for the lower tail used in [9] (and also in the **R** implementation) is just $g_l(0) = g_u(0)$. Our approximation is more accurate and provides upper and lower bounds.

For the upper tail, the iteration functions are the same as before, but with p and q interchanged (by using (1.3)). The bounds are then given for $1 - x$.

3 Numerical testing

In this section we illustrate the use of the different methods with numerical examples which will help in deciding how to combine the methods in order to obtain fast reliable algorithms (which are described in section 4).

3.1 Initial values obtained with the asymptotic approximations

In Tables 1 and 2 we show examples of the accuracy obtained in the computation of $|I_x(p,q) - \alpha| / \alpha$ with x the values provided by the asymptotic approximations (before iterating the SNM). [1]

The asymptotic methods provide relatively good initial values even for quite small values of p and q: using 10^7 random points in the region $(p, q, \alpha) \in (0.5,\, 1.5) \times (0.7,\, 1.5) \times (0,\, 1)$ we have tested that a relative accuracy better than 0.06 was obtained when computing $|I_x(p,q) - \alpha| / \alpha$ with x, the asymptotic approximations obtained using the error function. With these initial values, not more than two iterations of the SNM are needed to obtain an accuracy better than $5.0\,10^{-13}$. The function $I_x(p,q)$ is computed in the iterations of the SNM by using a continued fraction representation

$$I_x(p,q) = \frac{x^p(1-x)^q}{pB(p,q)} \left(\frac{1}{1+} \frac{d_1}{1+} \frac{d_2}{1+} \frac{d_3}{1+} \ldots \right), \tag{3.1}$$

where

[1] In Table 1 the accuracy $5.6\,10^{-16}$ corresponds to the case $I_x(3,3) = 0.5$: because of the symmetry, x should be 0.5 (exact), and η defined in (2.10) becomes 0, the same as η_0 in (2.12). This explains why that result in the table becomes so small.

α	$p=4$	$p=3$	$p=2$
10^{-6}	$6.3\,10^{-4}$	$1.6\,10^{-3}$	$1.8\,10^{-3}$
10^{-3}	$3.2\,10^{-4}$	$1.6\,10^{-3}$	$4.5\,10^{-3}$
0.1	$2.7\,10^{-4}$	$4.0\,10^{-4}$	$1.9\,10^{-3}$
0.3	$2.9\,10^{-5}$	$3.9\,10^{-6}$	$5.9\,10^{-5}$
0.5	$2.9\,10^{-5}$	$5.6\,10^{-16}$	$2.9\,10^{-5}$
0.7	$2.6\,10^{-5}$	$1.7\,10^{-6}$	$1.2\,10^{-5}$
0.9	$2.2\,10^{-4}$	$4.5\,10^{-5}$	$2.9\,10^{-5}$
0.999	$4.5\,10^{-6}$	$1.6\,10^{-6}$	$3.2\,10^{-7}$
0.99999	$2.9\,10^{-8}$	$1.8\,10^{-8}$	$6.2\,10^{-9}$

Table 1. Relative errors $|I_x(p,q)-\alpha|/\alpha$ for $r=p+q=6$ using the estimates provided by the asymptotic inversion method with the error function.

$$\begin{aligned} d_{2m} &= \frac{m(q-m)x}{(p+2m-1)(p+2m)}, \\ d_{2m+1} &= -\frac{(p+m)(p+q+m)x}{(p+2m)(p+2m+1)}. \end{aligned} \tag{3.2}$$

For the computation of the Beta function $B(p,q)$, it is convenient, in particular when p and q are large, to use the following expression in terms of the scaled gamma function $\Gamma^*(x)$:

$$B(p,q) = \sqrt{2\pi}\sqrt{\frac{1}{p}+\frac{1}{q}}\left(\frac{p^{\frac{p}{p+q}}q^{\frac{q}{p+q}}}{p+q}\right)^{p+q}\frac{\Gamma^*(p)\Gamma^*(q)}{\Gamma^*(p+q)}, \tag{3.3}$$

where $\Gamma^*(x)$ is defined as

$$\Gamma^*(x) = \frac{\Gamma(x)}{\sqrt{2\pi/x}\,x^x e^{-x}}, \quad x>0. \tag{3.4}$$

The function $\Gamma^*(x)$ is computed using the function **gamstar** included in a previous package developed by the authors [8].

3.2 Initial values obtained in the tails of the distribution function

The interval estimations in the tails of the central beta distribution function by using the fixed point iterations of §2.3, can be also used for providing starting values of the SNM. This will be particularly useful for quite small values of the parameters p and q, where the asymptotic method cannot be

α	$\mu = 0.1$	$\mu = 0.5$	$\mu = 2$
10^{-6}	$8.1\,10^{-5}$	$2.2\,10^{-4}$	$4.0\,10^{-4}$
10^{-4}	$3.3\,10^{-4}$	$2.3\,10^{-5}$	$2.2\,10^{-4}$
0.1	$2.4\,10^{-4}$	$1.9\,10^{-4}$	$2.5\,10^{-5}$
0.3	$1.3\,10^{-4}$	$1.4\,10^{-4}$	$3.6\,10^{-5}$
0.5	$7.8\,10^{-5}$	$9.8\,10^{-5}$	$3.2\,10^{-5}$
0.7	$3.9\,10^{-5}$	$6.1\,10^{-5}$	$2.4\,10^{-5}$
0.9	$1.1\,10^{-5}$	$2.3\,10^{-5}$	$1.1\,10^{-5}$
0.999	$1.0\,10^{-7}$	$3.0\,10^{-7}$	$2.2\,10^{-7}$
0.99999	$1.0\,10^{-9}$	$3.2\,10^{-9}$	$2.9\,10^{-9}$

Table 2. Relative errors $|I_x(p,q) - \alpha|\,/\alpha$ for $p = 7$ and several values of $\mu = q/p$ using the estimates provided by the asymptotic inversion method with the incomplete gamma function.

applied. It is important to note that when the value of the parameters p or q are close to 0, the inversion of $I_x(p,q)$ becomes problematic in the lower (when $p \to 0$) or upper (when $q \to 0$) tail of the cumulative distribution function, because of the particular shape of the functions.

In Table 3 we show the relative errors $1 - x_l/x$ and $1 - x_u/x$ obtained with the lower and upper bounds, respectively, for the solution x of the equation (1.4) for small values of p, q and α. The bounds (computed in the examples using Maple) are obtained with just three iterations of the fixed point methods of §2.3. We have also tested that for small values of p and q, the bounds provide in all cases reasonable approximations for starting the SNM, no matter if the value of α is small. Besides, even for not so small values of p and q, the bounds provide very accurate estimations when α is very small. In some cases, these estimations could be even better than the estimations of the asymptotic method.

3.3 Performance of the SNM for small values of the parameters

We have tested that the scheme for the SNM, as described in §2.1 when the parameters p and q are both small, also provides a good uniform accuracy in the computation: using 10^7 random points in the region $(p, q, \alpha) \in (0.1,\, 0.5) \times (0.1,\, 0.7) \times (0,\, 1)$ we have tested that a relative accuracy better than $4.8\,10^{-13}$ was obtained when computing $|I_x(p,q) - \alpha|\,/\alpha$. The maximum number of iterations of the SNM was 3.

α		$p = 0.3$ $q = 0.4$	$p = 0.4$ $q = 0.3$
10^{-7}	(LB)	$-1.2\,10^{-22}$	$-5.9\,10^{-17}$
	(UB)	$7.8\,10^{-49}$	$1.1\,10^{-49}$
10^{-5}	(LB)	$-5.4\,10^{-16}$	$-5.9\,10^{-12}$
	(UB)	$1.3\,10^{-48}$	$5.6\,10^{-36}$
10^{-3}	(LB)	$-2.5\,10^{-9}$	$-5.9\,10^{-7}$
	(UB)	$8.5\,10^{-29}$	$5.6\,10^{-21}$

Table 3. Relative errors $1 - x_l/x$ and $1 - x_u/x$ obtained with the lower (LB) and upper (UB) bounds for the solution x of the equation (1.4).

3.4 Efficiency testing

As we have shown in §3.1, the asymptotic method provides very accurate initial values for starting the SNM even for small values of the parameters p and q. But apart from accuracy, an important feature of any computational scheme is also efficiency. So, we have compared whether the combined use of the asymptotic approximations plus iterations of the SNM is more efficient or not than the sole use of the SNM. In Table 4 we show CPU times spent by 20000 runs of the inversion algorithm for different values of p, q and α using three methods of computation: a) asymptotic inversion method using the error function for estimating the initial value plus iterations of the SNM; b) asymptotic inversion method using the gamma function for estimating the initial value plus iteration of the SNM; c) iterations of the SNM with starting values obtained as discussed in §2.1. In all cases the SNM is iterated until the solution of equation (1.4) is obtained with an accuracy near full double precision.

The results in Table 4 and additional testing for other parameter values, indicate that the sole use of the SNM is efficient in all cases for the inversion of the cumulative central beta distribution, but specially when the values of the parameter α are neither very small nor near to 1.

A different scenario arises when the solution of equation (1.4) is computed with an aimed accuracy better than 10^{-8} (single precision). In this case, just using the approximations provided by the asymptotic expansions will be enough to obtain such an accuracy for a wide range of parameters.

α	Method	$p=4$ $q=3$	$p=50$, $q=60$	$p=100$, $q=80$	$p=150$, $q=1.0$	$p=300$, $q=400$
10^{-6}	M1	0.22	0.29	0.22	0.4	0.20
	M2	0.31	0.39	0.31	0.22	0.33
	M3	0.32	0.34	0.34	0.39	0.38
10^{-4}	M1	0.22	0.23	0.19	0.36	0.22
	M2	0.34	0.34	0.31	0.33	0.33
	M3	0.25	0.27	0.28	0.27	0.33
0.3	M1	0.19	0.16	0.17	0.20	0.17
	M2	0.31	0.28	0.30	0.19	0.30
	M3	0.11	0.19	0.20	0.19	0.17
0.7	M1	0.20	0.17	0.16	0.23	0.18
	M2	0.33	0.28	0.30	0.23	0.30
	M3	0.16	0.19	0.19	0.16	0.19
0.999	M1	0.17	0.17	0.17	0.14	0.16
	M2	0.25	0.28	0.28	0.16	0.33
	M3	0.25	0.27	0.28	0.16	0.27

Table 4. CPU times (in seconds) for 20000 runs of the inversion algorithm using different methods. M1: Asymptotic inversion method using the error function +SNM; M2: Asymptotic inversion method using the gamma function +SNM; M3: SNM. The SNM is iterated until the solution of equation (1.4) is obtained with an accuracy near full double precision.

4 Proposed algorithms

Based on the previous numerical experiments of §3, we conclude that if the precision required is not very high, the initial approximations given by asymptotics or by the tail estimations could be sufficient in a large range of the parameters. However, for higher precision the use of the SNM must be prevalent.

This leads us to suggest two different schemes for computing the solution x of the equation (1.4).

SCHEME 1. Algorithm for the inversion of the cumulative central beta distribution with an accuracy near double precision:

If $\alpha \leq 0.01$

For $p < 0.3$, use the upper bound of §2.3 as solution of the equation (1.4).

For $0.3 < p < 1$, use the SNM as described in §2.1 using as starting values the bounds of §2.3.

For $1 < p < 30$ and $q < 1$, use the SNM, using as starting values the bounds of §2.3.

For $p > 30$ and $q < 0.5$, use the SNM, using as starting values the bounds of §2.3.

For $p > 30$, $0.5 < q < 5$: a) if $\alpha > 0.0001$ use the SNM, using as starting values the approximation provided by the uniform asymptotic expansion in terms of the gamma function; b) if $\alpha < 0.0001$ use the SNM, using as starting values the bounds of §2.3.

In other cases, use the SNM as described, using as starting values the approximations provided by the uniform asymptotic expansion in terms of the error function in other cases.

When $0.01 < \alpha \leq 0.5$

For $1 < q < 5$ and $p > 50$, use the SNM, using as starting values the approximations provided by the uniform asymptotic expansion in terms of the incomplete gamma function.

For $p > 30$ and $q > 30$, use the SNM, using as starting values the approximations provided by the uniform asymptotic expansion in terms of the error function.

In other cases, use the SNM as described in §2.1.

For $0.5 < \alpha < 1$, use the relation (1.3) and apply the previous steps to solve $1 - x$ in $I_{1-x}(q, p) = 1 - \alpha$.

SCHEME 2. Algorithm for the inversion of the cumulative central beta distribution with an accuracy near single precision:

If $\alpha \leq 0.01$

For $p < 0.5$, use the upper bound of §2.3 as solution of the equation (1.4).

For $0.5 < p < 1$, use the SNM as described in §2.1 using as starting values the bounds of §2.3.

For $1 < p < 30$ and $q < 1$, use the SNM, using as starting values the bounds of §2.3.

For $p > 30$ and $q < 0.5$, use the SNM, using as starting values the bounds of §2.3.

For $p > 30$, $0.5 < q < 5$: a) if $\alpha > 0.0001$ use the SNM, using as starting values the approximation provided by the uniform asymptotic expansion in terms of the gamma function; b) if $\alpha < 0.0001$ use the SNM, using as starting values the bounds of §2.3.

In other cases, use the SNM as described in §2.1 using as starting values the approximations provided by the uniform asymptotic expansion in terms of the error function in other cases.

When $0.01 < \alpha \leq 0.5$

For $1 < q < 3$, $p > 160$ and $\alpha > 0.1$, use the approximation provided by the uniform asymptotic expansion in terms of the incomplete gamma function as solution of the equation (1.4).

For $p > 30$ and $q > 30$, use the approximation provided by the uniform asymptotic expansion in terms of the error function as solution of the equation (1.4).

In other cases, use the SNM as described in §2.1.

For $0.5 < \alpha < 1$, use the relation (1.3) and apply the previous steps to solve $1 - x$ in $I_{1-x}(q, p) = 1 - \alpha$.

5 Conclusions

We have presented methods for the inversion of cumulative beta distributions which improve previously existing methods. We have described how

the Schwarzian-Newton method provides a standalone method with certified convergence which is in itself an efficient method, even without accurate initial estimations (except at the tails). In addition, we have discussed how to improve the efficiency by estimating the quantile function using asymptotics for large p and/or q and by considering sharp upper and lower bounds for the tails. These initial estimations are considerably more accurate than the simple approximations used in some standard mathematical software packages (like **R**) and, combined with the fourth order SNM, provide efficient and reliable algorithms for the inversion of cumulative beta distributions.

6 Acknowledgements

The authors acknowledge financial support from *Ministerio de Economía y Competitividad*, project MTM2012-34787. NMT thanks CWI, Amsterdam, for scientific support.

References

[1] R.W. Abernathy and R.P. Smith. Algorithm 724: Program to Calculate F-Percentiles. *ACM Trans Math Soft*, 19(4):481–483, 1993.

[2] K. J. Berry, P. W. Mielke, and G. W. Cran. Algorithm as r83: A remark on algorithm as 109: Inverse of the incomplete beta function ratio. *Appl. Statist.*, 39(2):309–310, 1990.

[3] K. J. Berry, P. W. Mielke, and G. W. Cran. Correction to algorithm as r83-a remark on algorithm as 109: Inverse of the incomplete beta function ratio. *Appl. Statist.*, 40(1):236, 1991.

[4] G. W. Cran, K. J. Martin, and G. E. Thomas. Remark as r19 and algorithm as 109: A remark on algorithms: As 63: The incomplete beta integral as 64: Inverse of the incomplete beta function ratio. *Appl. Statist.*, 26(1):111–114, 1977.

[5] A.R. Didonato and A.H. Morris. Algorithm 708: Significant digit computation of the incomplete beta function ratios. *ACM Trans. Math. Softw.*, 18(3):360–373, 1992.

[6] A. Gil, J. Segura, and N. M. Temme. *Numerical Methods for Special Functions*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

[7] A. Gil, N. M. Temme, and J. Segura. Efficient and accurate algorithms for the computation and inversion of the incomplete gamma function ratios. *SIAM J. Sci. Comput.*, 34(6):A2965–A2981, 2012.

[8] A. Gil, N. M. Temme, and J. Segura. Gammachi: a package for the inversion and computation of the gamma and chi-square cumulative distribution functions (central and noncentral). *Comput. Phys. Commun.*, 191:132–139, 2015.

[9] K. L. Majumder and G. P. Bhattacharjee. Algorithm as 64: Inverse of the incomplete beta function ratio. *Appl. Statist.*, 2(3):411–414, 1973.

[10] J. Segura. The Schwarzian-Newton method for solving nonlinear equations, with applications. *Math. Comput. (to appear).*

[11] Javier Segura. Sharp bounds for cumulative distribution functions. *J. Math. Anal. Appl.*, 436(2):748–763, 2016.

[12] N. M. Temme. Asymptotic inversion of the incomplete beta function. *J. Comput. Appl. Math.*, 41(1-2):145–157, 1992.

[13] N. M. Temme. *Asymptotic methods for integrals.* World Scientific, Singapore, 2014. Series in Analysis, Vol. 6.