Algorithms for the computation and inversion of cumulative gamma distributions

Javier Segura

Departamento de Matemáticas, Estadística y Computación.
Universidad de Cantabria, Spain

In collaboration with Amparo Gil and Nico M. Temme.

Cumulative γ and χ^2 distributions

The cumulative central gamma distribution is given by the incomplete gamma function ratios

$$P_\mu(y) = \frac{\gamma(\mu, x)}{\Gamma(\mu)} = \frac{1}{\Gamma(\mu)} \int_0^y t^{\mu-1} e^{-t} dt,$$

The non-central γ distribution can be defined as

$$P_\mu(x, y) = e^{-x} \sum_{n=0}^{\infty} \frac{x^n}{n!} P_{\mu+n}(y),$$

From the Maclaurin series for modified Bessel functions $I_\mu(z)$ we obtain:

$$P_\mu(x, y) = x^{\frac{1}{2}(1-\mu)} \int_0^y t^{\frac{1}{2}(\mu-1)} e^{-t-x} I_{\mu-1}\left(2\sqrt{xt}\right) dt.$$

Observe that $P_\mu(0, y) = 0$, $P_\mu(0, +\infty) = 1$.
$P_\mu(0, y) = P_\mu(y)$.

The complementary distributions are obtained by changing the interval of integration from $[0, y]$, to $[y, +\infty)$. We have

$$Q_\mu(x, y) = x^{\frac{1}{2}(1-\mu)} \int_y^{+\infty} t^{\frac{1}{2}(\mu-1)} e^{-t-x} I_{\mu-1}\left(2\sqrt{xt}\right)$$

and

$$P_\mu(x, y) + Q_\mu(x, y) = 1.$$

The Q-function is also called generalized Marcum Q-function.

Similarly as before,

$$Q_\mu(y) = \frac{\Gamma(\mu, y)}{\Gamma(\mu)} = \frac{1}{\Gamma(\mu)} \int_y^{+\infty} t^{\mu-1} e^{-t} dt$$

and

$$Q_\mu(x, y) = e^{-x} \sum_{n=0}^{\infty} \frac{x^n}{n!} Q_{\mu+n}(y).$$

$P_\mu(x, y)$ is a cumulative distrubition function (CDF) with probability density function (PDF)

$$f_\mu(x,t) = \frac{1}{\Gamma(\mu)} x^{\frac{1}{2}(1-\mu)} t^{\frac{1}{2}(\mu-1)} e^{-t-x} I_{\mu-1}\left(2\sqrt{xt}\right),$$

that is,

$$P_\mu(x,y) = \int_0^y f_\mu(x,t)dt$$

For a fixed values of x, μ and p, denote as $F_\mu(x,p)$ the inverse with respect to y:

$$P_\mu(x, F_\mu(x,p)) = p.$$

Random number generation:

If U is a random uniform variable in $[0,1]$ then $Y = F_\mu(x,U)$ is a random variable with PDF as before.

A different way to generate random samples with a given PDF is acceptance-rejection method (the CDF is needed in this case)

An analysis of the situation revealed that:

1. Algorithms existed both for the computation and inversion (with respect to y) of the CDF for the central case, but that asymptotics could improve both algorithms.
2. No algorithms were available for the computation of the CDFs for the noncentral case for real μ, and the available software had a limited range of validity (and also some inaccuracies, as we will see).
3. Algorithms for the inversion in the noncentral case also had a limited range of validity (the inversion with respect to x is also important in applications). No published software was available.

Computation of the central distribution

Recall the definitions

$$P_a(y) = \frac{1}{\Gamma(a)} \int_0^y t^{a-1} e^{-t} dt, Q_a(y) = \frac{1}{\Gamma(a)} \int_y^{+\infty} t^{a-1} e^{-t} dt$$

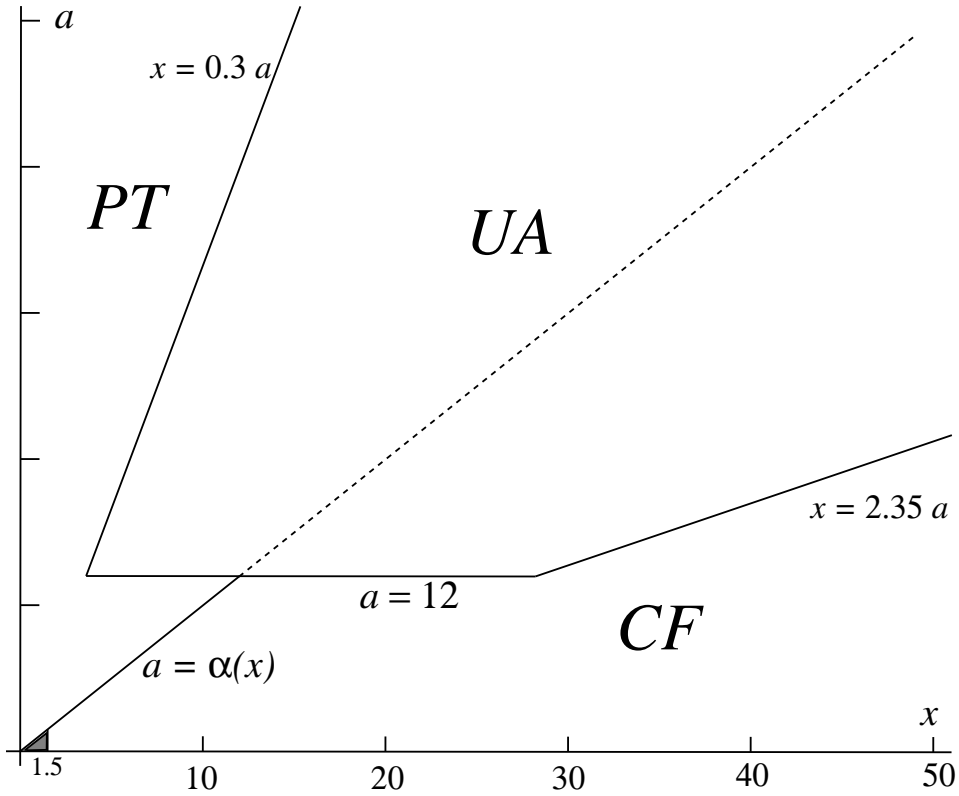
Because $P_a(y) + Q_a(y) = 1$ we only need to compute one function. We compute the smallest of the two.

For large values of a, x we have a transition at $a \sim x$, with

$$P_a(y) \lesssim \tfrac{1}{2} \quad \text{when} \quad a \gtrsim y,$$

$$Q_a(y) \lesssim \tfrac{1}{2} \quad \text{when} \quad a \lesssim y.$$

Accordingly, the methods of computation are divided in two zones, with several methods of computation in each one.



PT: Taylor series for P.
QT: Taylor series for Q (small triangle).
UA: Uniform asymptotic expansions.
CF: continued fraction for the Q.

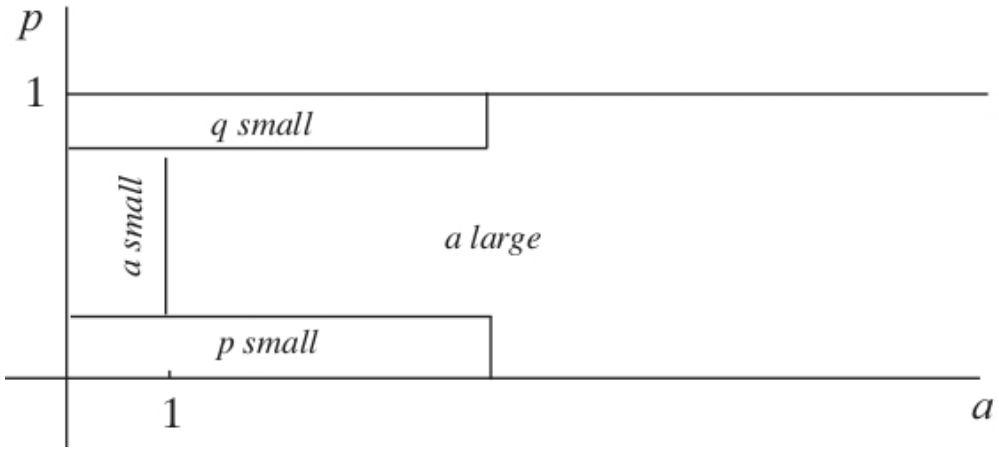
Inversion of the central distribution

As commented, the inversion is also needed in applications.
For fixed a, we invert $P_a(y) = p$ or, equivalently, $Q_a(y) = q$.

Our approach:

1. Invert $P_a(y)$ ($Q_a(y)$) if $p < q$ ($p > q$)
2. Use the existent approximation methods (PT, Poincaré asymptotics for Q, UA) to find starting values.
3. Apply higher order Newton methods from the resulting starting values.

The different type of starting values are chosen according to the next figure.



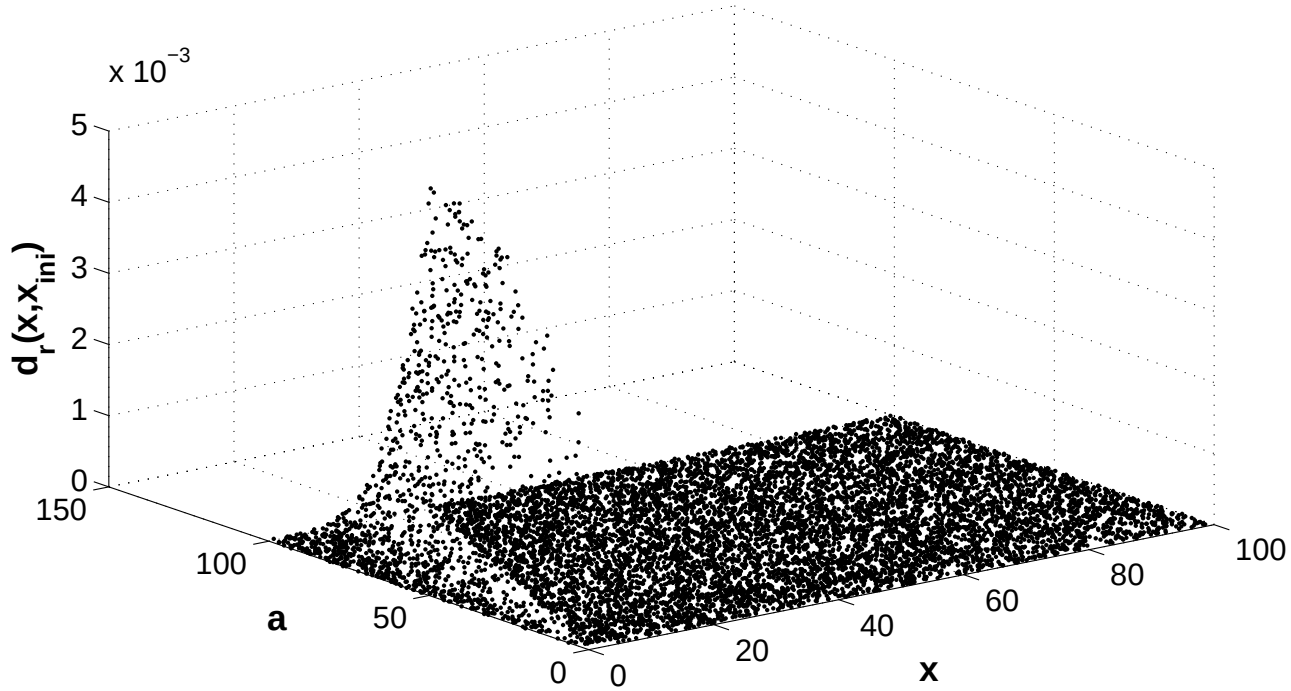
Starting values (an example):
For small p, we use PT to write

$$x = r + \sum_{n=2}^{\infty} c_k r^k,$$

where $r = (p\Gamma(1+a))^{1/a}$ and by expanding, the first few coefficients are

$$c_2 = \frac{1}{a+1},$$
$$c_3 = \frac{3a+5}{2(a+1)^2(a+2)},$$
$$c_4 = \frac{8a^2+33a+31}{3(a+1)^3(a+2)(a+3)},$$
$$c_5 = \frac{125a^4+1179a^3+3971a^2+5661a+2888}{24(1+a)^4(a+2)^2(a+3)(a+4)}.$$

The accuracy of the starting values is shown in this figure



2-3 iterations (in the worst cases) with a fourth order Newton method give a rel. accuracy $\sim 10^{-13}$

A. Gil, J. Segura, N.M. Temme. SIAM J Sci Comput 34(6) (2012) A2965-A2981.

W. Gautschi. ACM Trans. Math. Software 5 (1979) 466–481.

A. R. DiDonato, A. H. Morris. ACM Trans. Math. Software 12 (1986) 377–393.

Computation of the non-central distribution

The generalized Marcum Q-function is the **non-central cumulative χ^2 distribution**, up to elementary redefinition of the variables. We recall that

$$Q_\mu(x,y) = x^{\frac{1}{2}(1-\mu)} \int_y^{+\infty} t^{\frac{1}{2}(\mu-1)} e^{-t-x} I_{\mu-1}\left(2\sqrt{xt}\right)\, dt,$$

where $\mu > 0$ and $I_\mu(z)$ is the modified Bessel function.

The complementary function such that $P_\mu(x,y) + Q_\mu(x,y) = 1$:

$$P_\mu(x,y) = x^{\frac{1}{2}(1-\mu)} \int_0^{y} t^{\frac{1}{2}(\mu-1)} e^{-t-x} I_{\mu-1}\left(2\sqrt{xt}\right)\, dt.$$

Particular values are

$$Q_\mu(x,0) = 1, \qquad Q_\mu(x,+\infty) = 0,$$

$$Q_\mu(0,y) = Q_\mu(y), \qquad Q_\mu(+\infty,y) = 1,$$

$$Q_{+\infty}(x,y) = 1.$$

As for incomplete gamma functions, we compute the smallest of the two functions. Asymptotic analysis gives that for large values of μ, x, y, we have a transition at $y \sim x + \mu$, with

$$P_\mu(x,y) \lesssim Q_\mu(x,y) \quad \text{when} \quad y \lesssim x+\mu,$$

$$Q_\mu(x,y) \lesssim P_\mu(x,y) \quad \text{when} \quad y \gtrsim x+\mu.$$

Ingredients in the computation:

1. Series in terms of incomplete gamma functions.
2. Recurrence relations.
3. Asymptotic expansions for large μ in terms of the error function (both for P and Q).
4. Asymptotic expansions for large xy.
5. Quadrature methods.

We give some details of these methods.

Series in incomplete gamma functions

We can consider the series expansions

$$P_\mu(x,y) = e^{-x}\sum_{n=0}^{\infty}\frac{x^n}{n!}P_{\mu+n}(y),$$
$$Q_\mu(x,y) = e^{-x}\sum_{n=0}^{\infty}\frac{x^n}{n!}Q_{\mu+n}(y).$$

in terms of the incomplete gamma function ratios (which we can compute). Recurrences can be used to compute rapidly the series. We have

$$Q_{\mu+1}(y) = Q_\mu(y) + \frac{y^\mu e^{-y}}{\Gamma(\mu+1)},$$
$$P_{\mu+1}(y) = P_\mu(y) - \frac{y^\mu e^{-y}}{\Gamma(\mu+1)},$$

stable for $Q_\mu(y)$ in the forward direction, and for $P_\mu(y)$ in the backward direction. Equivalently, we have $\mu Q_{\mu+1}(y) - (\mu+y)Q_\mu(y) + yQ_{\mu-1}(y) = 0$.

The series

$$Q_\mu(x,y) = e^{-x} \sum_{n=0}^{\infty} \frac{x^n}{n!} Q_{\mu+n}(y)$$

can be computed from two values $Q_\mu(y)$ and $Q_{\mu+1}(y)$ and forward recursion with

$$Q_{\mu+1}(y) = \left(1 + \frac{y}{\mu}\right) Q_\mu(y) - \frac{y}{\mu} Q_{\mu-1}(y)$$

For the other series, we write

$$P_\mu(x,y) \simeq e^{-x} P_\mu(y) \sum_{n=0}^{n_0} \frac{x^n}{n!} \frac{P_{\mu+n}(y)}{P_\mu(y)},$$

estimate the value n_0 which gives sufficient accuracy and compute using the backward recursion

$$P_{\mu-1}(y) = -\frac{\mu}{y} P_{\mu+1}(y) + \left(1 + \frac{\mu}{y}\right) P_\mu(y)$$

Recurrence relations

Integration by parts gives the following recurrences

$$Q_{\mu+1}(x,y) = Q_\mu(x,y) + \left(\frac{y}{x}\right)^{\mu/2} e^{-x-y} I_\mu(2\sqrt{xy}),$$

$$P_{\mu+1}(x,y) = P_\mu(x,y) - \left(\frac{y}{x}\right)^{\mu/2} e^{-x-y} I_\mu(2\sqrt{xy}),$$

It is possible to eliminate the Bessel function and obtain a homogeneous recurrence relation.

$$xQ_{\mu+2}(x,y) = (x-\mu)Q_{\mu+1}(x,y) + (y+\mu)Q_\mu(x,y) - yQ_{\mu-1}(x,y),$$

and $P_\mu(x,y)$ satisfies the same relation, but its computation with this recurrence is badly conditioned (it is subdominant, but not minimal)
A better possibility is:

$$y_{\mu+1} - (1+c_\mu)y_\mu + c_\mu y_{\mu-1} = 0, \quad c_\mu = \sqrt{\frac{y}{x}}\frac{I_\mu\left(2\sqrt{xy}\right)}{I_{\mu-1}\left(2\sqrt{xy}\right)}.$$

P is minimal and Q is dominant. Pincherle's theorem gives:

$$\frac{P_\mu(x,y)}{P_{\mu-1}(x,y)} = \frac{c_\mu}{1+c_\mu-} \frac{c_{\mu+1}}{1+c_{\mu+1}-} \cdots$$

Asymptotic expansions for μ large

We start from

$$Q_{\mu+1}(\mu x, \mu y) = \frac{\mu e^{-\mu x}}{(2x)^{\mu+1}} \int_{\xi}^{\infty} z e^{-\mu\phi(z)} e^{-\mu\eta(z)} I_{\mu}(\mu z)\, dz,$$

where

$$\phi(z) = -\ln z + \frac{1}{4x} z^2 - \eta(z),\ \eta(z) = \sqrt{1+z^2} + \log \frac{z}{1+\sqrt{1+z^2}},\ \xi = 2\sqrt{xy}$$

The saddle point follows from the equation $\phi'(z) = 0$. It follows that the positive saddle point z_0 is given by

$$z_0 = 2\sqrt{x(1+x)}. \tag{1}$$

The transition line in the scaled variables is $y = x + 1$.

The saddle point coalesces with the end point of integration as $y \to x + 1$. Bleinstein's method is a good choice (we omit details).

Q is computed for $y > x + 1$ (in the unscaled variables $y > x + \mu$).
For P analogous expansions can be worked out ($y < x + \mu$)

$$Q_\mu(\mu x, \mu y) \sim \tfrac{1}{2}\mathrm{erfc}\left(-\zeta\sqrt{\mu/2}\right) + \sqrt{\frac{\mu}{2\pi}}\sum_{k=1}^{\infty} B_k - e^{-\frac{1}{2}\mu\zeta^2} e^{-\mu\eta(\xi)} I_\mu(\mu\xi).$$

$$B_k = \sum_{j=0}^{k} \frac{f_{j,k-j}\Psi_j(\zeta)}{\mu^{k-j}}$$

$$\Psi_j(\zeta) = \left(\frac{2}{\mu}\right)^{(j+1)/2} \int_{-\zeta\sqrt{\mu/2}}^{\infty} e^{-s^2} s^j \, ds.$$

which can be written in terms of incomplete gamma functions.

$$\zeta = sign(x+1-y)\sqrt{2\left(\phi(\xi) - \phi(z_0)\right)}.$$

$$f_k(w) = \frac{z}{2x}\frac{u_k(t)}{(1+z^2)^{\frac{1}{4}}}\frac{dz}{dw} = \sum_{j=0}^{\infty} f_{jk}(w-\zeta)^j, \; t = 1/\sqrt{1+z^2}$$

$$\phi(z) - \phi(\xi) = \tfrac{1}{2}w^2 - \zeta w, \xi = 2\sqrt{xy}$$

$$u_0(t) = 1, \qquad u_1(t) = \frac{3t - 5t^3}{24}, \qquad u_2(t) = \frac{81t^2 - 462t^4 + 385t^6}{1152},$$

and other coefficients can be obtained by applying the formula

$$u_{k+1}(t) = \tfrac{1}{2}t^2(1-t^2)u_k'(t) + \tfrac{1}{8}\int_0^t (1-5s^2)u_k(s)\,ds, \quad k = 0, 1, 2, \ldots.$$

Numerical quadrature

We start from the contour integral representation for the function $Q_\mu(x, y)$

$$Q_\mu(x, y) = \frac{e^{-x-y}}{2\pi i} \int_{\mathcal{L}_\mathcal{Q}} \frac{e^{x/s+ys}}{1-s} \frac{ds}{s^\mu},$$

where $\mathcal{L}_\mathcal{Q}$ is a vertical line that cuts the real axis in a point s_0, with $0 < s_0 < 1$. Introducing scaled variables x, y and integrating along the path of steepest descent, we arrive to:

$$Q_\mu(\mu x, \mu y) = \frac{e^{-\frac{1}{2}\mu\zeta^2}}{2\pi} \int_{-\pi}^{\pi} e^{-\mu\psi(\theta)} f(\theta)\, d\theta,$$

for specific ζ, ψ, f.

The integrand is analytic and vanishing with all derivatives at $\pm\pi$

The trapezoidal rule is very efficient for this type of integrals.

The methods are combined as follows:

Let $\xi = 2\sqrt{xy}$ and

$$f_1(x,\mu) = x + \mu - \sqrt{4x + 2\mu}, \quad f_2(x,\mu) = x + \mu + \sqrt{4x + 2\mu}. \tag{2}$$

Then the scheme is as follows:

1. If $x < 30$, then compute the series expansion.
2. If $\xi > 30$ and $\mu^2 < 2\xi$, then compute the asymptotic expansion for large ξ.
3. If $f_1(x,\mu) < y < f_2(x,\mu)$ and $\mu < 135$, then compute the Marcum functions using three-term recurrence relations.
4. If $f_1(x,\mu) < y < f_2(x,\mu)$ and $\mu \geq 135$, then use the asymptotic expansion for μ large.
5. In other case: compute the integral representation.

Our implementation: A. Gil, J. Segura, N.M. Temme, *Algorithm 939: Computation of the Marcum Q-function* ACM Trans. Math. Soft. 40(3) (2014).

An accuracy $\sim 10^{-12}$ is obtained in the parameter region $(x, y, \mu) \in [0,\, A] \times [0,\, A] \times [1,\, A]$, for $A = 200$. For larger parameters the accuracy decreases a little (close to $5\,10^{-11}$ for $A = 10^5$).

Previous work includes:

C.W. Helstrom. IEEE Trans. Inf. Theory (1992)

D.A. Shnidman. IEEE Trans. Inf. Theory (1989)

But no verified public software was available until algorithm 939.

Previously existing software reduces to MATLAB & Mathematica. Are they reliable?

Mathematica is usually much slower that our codes. The advantage is that one can use as many digits as needed, but it is not always clear how many digits are needed.

If the demanded number of digits is not high enough, Mathematica may fail.
First example: $P_1(800, 200)$**.**

Computing with $D = 16$, Mathematica 8 gives the negative result $-7.170326438841136\,10^{-42}$.

When D is not declared, Mathematica gives $1.94499\,10^{-89}$; these digits are correct.

With $D \leq 64$, Mathematica still returns a negative value, but the order of magnitude becomes better.

With $D = 65$, Mathematica gives three correct digits for this example.

Second example: $Q_1(480.5, 200)$**.**

Computing with $D = 16$, Mathematica 8 gives the result $7.607098470200109\,10^{20}$.

When D is not declared, Mathematica gives 1, which is the correct limiting value when $x \gg y$.

Declaring $D = 37$ is needed to obtain three correct digits.

We conclude about Mathematica:

- Mathematica 8 can always give the correct result when D is high enough.
- Asking for high accuracy D does not always give the correct D digits.

What about MATLAB?

MATLAB is based on Shnidman algorithm and has the following limitations:

- Only integer values of μ.
- Only Marcum $Q-$function, not $P-$function.
- Slower performances than our algorithms.

Serious errors in MATLAB!

- Especially near the transition line $y = x + \mu$.
- The plot of the increasing function $Q_2(x, 200)$ as a function of x shows very rapid oscillations for $0 < x < 2400$.
- The plot of the decreasing function $Q_{800}(1, y)$ as a function of y shows several abrupt changes in the interval $[750, 850]$, with a steep jump close to $y = 800$. Results are meaningless for $800 < y < 1100$.
- And many more errors were found.

Inversion of the noncentral distribution

The problem is inverting with respect to x or y the equations

$$Q_\mu(x,y) = q, \quad P_\mu(x,y) = p.$$

In statistics,

- the inversion of $Q_\mu(x,y)$ with respect to x corresponds to the problem of inverting the distribution function with respect to the noncentrality parameter given the upper tail probability.
- The inversion of $P_\mu(x,y)$ with respect to y with fixed x corresponds to the problem of computing the p-quantiles of the distribution function. This is related to the generation of random numbers corresponding to a noncentral gamma distribution

Some ingredients of the methods:

- For large μ we use an asymptotic representation; essential steps in the inversion algorithm are based on the inversion of the complementary error function.
- We use monotonicity and convexity properties of these functions for finding initial values for reliable Newton or secant methods to invert the functions.
- We set a more modest relative accuracy goal $\sim 10^{-10}$. Two secant steps are usually necessary (and 4 steps in the worst cases).

Our methods are described in:
A. Gil, J. Segura, N.M. Temme. The asymptotic and numerical inversion of the Marcum *Q*-function. Studies in Applied Mathematics (to appear).

An implementation of these ideas can be found in:
A. Gil, J. Segura, N.M. Temme. GammaCHI: a package for the inversion and computation of the gamma and chi-square cumulative distribution functions (central and noncentral). Submitted to Computer Physics Communications.

A previous reference for inversion:

C.W. Helstrom. IEEE Trans. Aerosp. Electron. Syst. (1998)

Next steps:

1. About direct computation: should we consider a generalization of Marcum-Q? The Nuttal Q-function:

$$Q_{\eta,\mu}(x,y) = x^{\frac{1}{2}(1-\mu)} \int_y^{+\infty} t^{\eta+\frac{1}{2}(\mu-1)} e^{-t-x} I_{\mu-1}\left(2\sqrt{xt}\right)\, dt,$$

$$Q_{0,\mu}(x,y) = Q_\mu(x,y)$$

2. Other important cumulative distribution functions: the beta distribution function.

Thank you!