ELSEVIER Computational Statistics & Data Analysis 22 (1996) 345–350

COMPUTATIONAL STATISTICS & DATA ANALYSIS

On the computation of the distribution of the square of the sample multiple correlation coefficient

Cherng G. Ding

Institute of Management Science, National Chiao Tung University, 4F, 114 Chung-Hsiao W. Road, Section 1, Taipei, Taiwan, ROC

Abstract

Two computationally simple methods are proposed to evaluate the distribution function and the density of the square of the sample multiple correlation coefficient. No auxiliary routine is required. The accuracy of recursive computations can be effectively controlled. The distribution function and the density can be evaluated concurrently because their computing formulas are closely related. This property can enhance efficiency of Newton's method for computing the quantiles of the distribution. The corresponding algorithms are provided in a step-by-step form.

Keywords: Central beta distribution; Error bound; Gamma function; Multiple correlation coefficient; Newton's method; Noncentral beta distribution; Quantile; Series representation

1. Introduction

Let $X_1, X_2, \ldots, X_m$ have a multivariate normal distribution with mean vector μ and covariance matrix Σ, and R be the sample multiple correlation coefficient between X_1 and $X_2, \ldots, X_m$ based on a sample of size $N > m$. The density of $Y = R^2$ can be expressed as an infinite weighted sum of central beta densities as

0167-9473/96/$15.00
PII S0167-9473(96)00002-3

follows (see, e.g., Anderson, 1984, p. 145):

$$f(y; m, N, \rho^2) = \sum_{i=0}^{\infty} \frac{\Gamma^2[(N-1)/2+i](\rho^2)^i(1-\rho^2)^{(N-1)/2} y^{(m-1)/2+i-1}(1-y)^{(N-m-2)/2}}{\Gamma[(N-1)/2]i!\,\Gamma[(m-1)/2+i]\Gamma[(N-m)/2]}$$

$$= \sum_{i=0}^{\infty} q_i f(y; a+i, b), \tag{1}$$

where $0 \le y \le 1$, ρ is the population multiple correlation coefficient, $\Gamma(\alpha)$ is the gamma function, $a = (m-1)/2$, $b = (N-m)/2$, $q_i = (\Gamma(a+b+i)/\Gamma(a+b)i!)(\rho^2)^i(1-\rho^2)^{a+b}$, and $f(y; a, b)$ is the central beta density with shape parameters a and b. The distribution function of Y can, similarly, be expressed as an infinite weighted sum of central beta distribution functions. That is,

$$P(Y \le y) = F(y; m, N, \rho^2) = \sum_{i=0}^{\infty} q_i F(y; a+i, b), \tag{2}$$

where $F(y; a, b)$ is the central beta distribution function with shape parameters a and b. A recursive algorithm for evaluating $F(y; m, N, \rho^2)$ based on the series representation in (2) was given by Ding and Bargmann (1991a). The algorithm sums up the terms in (2) until the derived upper bound for the error of truncation is less than some predetermined accuracy. Auxiliary routines for evaluating $F(y; a, b)$ and the natural logarithm of the gamma function are required. Ding and Bargmann (1991b) also computed the quantile y_p such that $f(y_p; m, N, \rho^2) = p$ for given values of $m(>1)$, $N(>m)$, $\rho^2(0 \le \rho^2 \le 1)$, and $p(0 \le p \le 1)$ by using the Illinois method. In this paper, computationally simple methods are proposed to evaluate $F(y; m, N, \rho^2)$ (based on an alternative series representation) and $F(y; m, N, \rho^2)$ (based on the series representation in (1)). No auxiliary routine is required. The computational accuracy can be effectively controlled by using the error bounds obtained. Moreover, the recurrence formulas for computing $F(y; m, N, \rho^2)$ and $f(y; m, N, \rho^2)$ are closely related, and therefore $F(y; m, N, \rho^2)$ and $f(y; m, N, \rho^2)$ can be evaluated concurrently. This property can enhance efficiency of Newton's method for computing the quantile y_p. The corresponding algorithms are provided in a step-by-step form. Numerical methods and the resultant algorithms basically follow those given in Ding (1994) for computing the noncentral beta distribution, which is an infinite weighted sum of central beta distributions with Poisson weights.

2. Numerical methods

The numerical methods discussed in this section for evaluating $F(y; m, N, \rho^2)$ and $f(y; m, N, \rho^2)$ is for $0 < y < 1$. No computation is needed for $y = 0$ or $y = 1$ since $f(0; m, N, \rho^2) = f(1; m, N, \rho^2) = F(0; m, N, \rho^2) = 0$, and $F(1; m, N, \rho^2) = 1$.

A recursive formula for evaluating $F(y; a+i, b)$ in (2) was given by Ding (1994) as follows:

$$F(y; a+i, b) = \sum_{k=i}^{\infty} \frac{(1-y)}{(a+b+k)} f(y; a+k+1, b) = \sum_{k=i}^{\infty} t_k, \quad i = 0, 1, \ldots, \tag{3}$$

where

$$t_0 = \frac{\Gamma(a+b)}{\Gamma(a+1)\Gamma(b)} y^a (1-y)^b,$$

$$t_i = t_{i-1} y(a+b+i-1)/(a+i), \quad i \geq 1. \tag{4}$$

It follows that $F(y; m, N, \rho^2)$ can be expressed by a new series in term of central beta densities:

$$\begin{aligned} F(y; m, N, \rho^2) &= \sum_{i=0}^{\infty} q_i \left(\sum_{k=i}^{\infty} t_k \right) \\ &= \sum_{i=0}^{\infty} \left(\sum_{k=0}^{i} q_k \right) t_i \\ &= \sum_{i=0}^{\infty} v_i t_i, \end{aligned} \tag{5}$$

where the terms are evaluated recursively by

$$\begin{aligned} &v_0 = q_0 = (1-\rho^2)^{a+b}, \\ &v_i = v_{i-1} + q_i, \quad q_i = q_{i-1}(a+b+i-1)(\rho^2)/i, \quad i \geq 1, \\ &t_i,\ i \geq 0, \quad \text{as in (4).} \end{aligned} \tag{6}$$

Since m and N are both integers, the gamma function in (4) is easy to evaluate in a sense that $\Gamma(\alpha) = (\alpha-1)!$ if α is an integer, and $\Gamma(\alpha) = (\alpha-1)\ldots\frac{1}{2}\sqrt{\pi}$ if α is a half-integer. $F(y; m, N, \rho^2)$ can be approximated by the finite sum $\sum_{i=0}^{n-1} v_i t_i$. Let EF_n denote the error of truncation. Using the facts that $\sum_{i=0}^{\infty} q_i = 1$ (see, e.g., Abramowitz and Stegun, 1965, p. 556), and $\sum_{i=n}^{\infty} t_i \leq t_{n-1} y(a+b+n-1)/[(a+n)-(a+b+n)y]$ if $a+n > (a+b+n)y$ (see Ding, 1994), we have, under the same condition,

$$EF_n = \sum_{i=n}^{\infty} v_i t_i \leq \sum_{i=n}^{\infty} t_i \leq t_{n-1} y(a+b+n-1)/[(a+n)-(a+b+n)y]. \tag{7}$$

The error bound given above is a decreasing function of n when $a+n > (a+b+n)y$. To evaluate $F(y; m, N, \rho^2)$, accumulate the terms in (5), computed recursively through (6), until the error bound is not greater than a specified accuracy ε. In fact, the evaluation of $F(y; m, N, \rho^2)$ requires no auxiliary routine.

The density $f(y; m, N, \rho^2)$ of Y expressed in (1) is another series in terms of central beta densities. Its terms can be evaluated recursively, and are related to

those of (5) as follows:

$$f(y; m, N, \rho^2) = \sum_{i=0}^{\infty} q_i f(y; a+i, b) = \sum_{i=0}^{\infty} q_i s_i, \tag{8}$$

where

$$s_0 = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1} = at_0/y/(1-y),$$

$$s_i = s_{i-1} y(a+b+i-1)/(a+i-1)$$

$$= t_{i-1}(a+b+i-1)/(1-y), \quad i \geq 1, \tag{9}$$

q_i and t_i, $i \geq 0$, are those in (6).

Let Ef_n be the truncation error at $i = n$ for the series in (8). Since the sequence $\{s_i\}(i \geq n)$ is decreasing when $a + n > (a + b + n)y$, we have, under the same condition,

$$Ef_n = \sum_{i=n}^{\infty} q_i s_i < \sum_{i=n}^{\infty} q_i s_n = s_n\left(1 - \sum_{i=0}^{n-1} q_i\right) = s_n(1 - v_{n-1}). \tag{10}$$

Likewise, the above error bound is a decreasing function of n, and is used to control the accuracy of the evaluation of $f(y; m, N, \rho^2)$. Again, the evaluation of $f(y; m, N, \rho^2)$ requires no auxiliary routine.

Let $G(y) = F(y; m, N, \rho^2) - p$. The quantile y_p is to be obtained by solving the equation $G(y) = 0$. Since $G(0) = -p$, $G(1) = 1 - p$, and G is strictly increasing in $[0, 1]$, the solution y_p of $G(y) = 0$ is unique. An efficient root-finding method is Newton's method, which requires the evaluations of both of $F(y; m, N, \rho^2)$ and $G'(y) = f(y; m, N, \rho^2)$. The process is to repeat computing (see, e.g., Kennedy and Gentle, 1980, pp. 72–73)

$$y_{j+1} = y_j - \frac{[F(y_j; m, N, \rho^2) - p]}{f(y_j; m, N, \rho^2)}, \quad j = 0, 1, \ldots, \tag{11}$$

until $|y_{j+1} - y_j| \leq \delta y_{j+1}$, where δ is a specified accuracy. It is obvious that $y_p = 0$ for $p = 0$ and $y_p = 1$ for $p = 1$. No computation is needed for these cases. For $0 < p < 1$, perform iterations (11) with the starting value $y_0 = 0.5$. For each iteration, $F(y_j; m, N, \rho^2)$ and $f(y_j; m, N, \rho^2)$ can be evaluated concurrently rather than independently because their computing formulas (5) and (8) are closely related. This property can greatly enhance the computational efficiency of Newton's method. To ensure the legality of the iterate y_{j+1}, use the same adjustments as in Ding (1994): if $y_{j+1} \leq 0$, replace it by $y_j/2$; if $y_{j+1} \geq 1$, replace it by $(y_j + 1)/2$. Note that the evaluations of $F(y_j; m, N, \rho^2)$ and $f(y_j; m, N, \rho^2)$ should be precise enough so that the accuracy of Newton's solution y_p can be ensured. Also, the number of iterations should be controlled.

3. Algorithms

According to the formulas discussed in Section 2, we provide three effective algorithms in a step-by-step form for evaluating $F(y; m, N, \rho^2), f(y; m, N, \rho^2)$, and the quantile y_p. The algorithm for computing y_p is particularly efficient.

Algorithm A. This algorithm computes the distribution function $F(y; m, N, \rho^2)$ of $Y = R^2$ for given values of $y(0 < y < 1)$, $m(> 1)$, $N(> m)$, and $\rho^2(0 \le \rho^2 \le 1)$.

A1 (Specify the accuracy). Set $EPS \leftarrow \varepsilon$ (some desired accuracy, e.g., 10^{-6}).

A2 (Set the constants). Set $a \leftarrow (m-1)/2$, $b \leftarrow (N-m)/2$.

A3 (Initialize). Set $n \leftarrow 1$, $t \leftarrow (\Gamma(a+b)/\Gamma(a+1)\Gamma(b))\, y^a(1-y)^b$, $q \leftarrow (1-\rho^2)^{a+b}$, $v \leftarrow q$, $CDF \leftarrow vt$.

A4 (Compare $a+n$, $(a+b+n)y$). If $a+n > (a+b+n)y$, go to step A6.

A5 (Update the term and then accumulate. Then increase n by 1). Set $q \leftarrow q(a+b+n-1)(\rho^2)/n$, $v \leftarrow v+q$, $t \leftarrow ty(a+b+n-1)/(a+n)$, $CDF \leftarrow CDF + vt$, $n \leftarrow n+1$, and return to step A4.

A6 (Find the error bound and check for convergence). Set $bound \leftarrow ty(a+b+n-1)/((a+n)-(a+b+n)y)$. If $bound \le EPS$, terminate the algorithm. (CDF is the answer.)

A7 (Update the term and then accumulate. Then increase n by 1). Set $q \leftarrow q(a+b+n-1)(\rho^2)/n$, $v \leftarrow v+q$, $t \leftarrow ty(a+b+n-1)/(a+n)$, $CDF \leftarrow CDF + vt$, $n \leftarrow n+1$, and return to step A6.

Algorithm B. This algorithm computes the density $f(y; m, N, \rho^2)$ of $Y = R^2$ for given values of $y(0 < y < 1)$, $m(> 1)$, $N(> m)$, and $\rho^2(0 \le \rho^2 \le 1)$.

B1 (Specify the accuracy). Set $EPS \leftarrow \varepsilon$ (some desired accuracy, e.g., 10^{-6}).

B2 (Set the constants). Set $a \leftarrow (m-1)/2$, $b \leftarrow (N-m)/2$.

B3 (Initialize). Set $n \leftarrow 1$, $s \leftarrow (\Gamma(a+b)/\Gamma(a)\Gamma(b))\, y^{a-1}(1-y)^{b-1}$, $q \leftarrow (1-\rho^2)^{a+b}$, $v \leftarrow q$, $PDF \leftarrow qs$.

B4 (Compare $a+n$, $(a+b+n)y$). If $a+n > (a+b+n)y$, go to step B6.

B5 (Update the term and then accumulate. Then increase n by 1). Set $q \leftarrow q(a+b+n-1)(\rho^2)/n$, $v \leftarrow v+q$, $s \leftarrow sy(a+b+n-1)/(a+n-1)$, $PDF \leftarrow PDF + qs$, $n \leftarrow n+1$, and return to step B4.

B6 (Find the error bound and check for convergence). Set $bound \leftarrow sy(a+b+n-1)(1-v)/(a+n-1)$. If $bound \le EPS$, terminate the algorithm. (PDF is the answer.)

B7 (Update the term and then accumulate. Then increase n by 1). Set $q \leftarrow q(a+b+n-1)(\rho^2)/n$, $v \leftarrow v+q$, $s \leftarrow sy(a+b+n-1)/(a+n-1)$, $PDF \leftarrow PDF + qs$, $n \leftarrow n+1$, and return to step B6.

Algorithm C. This algorithm computes the quantile y_p of the distribution of $Y = R^2$ such that $F(y_p; m, N, \rho^2) = p$ for given values of $m(> 1)$, $N(> m)$, $\rho^2(0 \le \rho^2 \le 1)$, and $p(0 < p < 1)$.

C1 (Specify the accuracy and the maximum number of Newton's iterations allowed). Set $EPS \leftarrow \varepsilon$ (some desired accuracy, e.g., 10^{-6}, for computing $F(y; m, N, \rho^2)$

and $f(y; m, N, \rho^2)$), $DELTA \leftarrow \delta$ (some desired accuracy, e.g., 10^{-4}, for computing y_p), $ITRMAX \leftarrow N_{\max}$ (an integer, e.g., 10, for controlling the number of Newton's iterations).

C2 (Set the constants). Set $a \leftarrow (m-1)/2$, $b \leftarrow (N-m)/2$, $coeff \leftarrow \Gamma(a+b)/\Gamma(a+1)\Gamma(b)$, $q_0 \leftarrow (1-\rho^2)^{a+b}$.

C3 (Loop on k, Newton's iteration). First initialize y by setting $y \leftarrow 0.5$. Then perform steps C4–C10 for $k = 1, 2, \ldots, ITRMAX$. (Steps C4–C10 constitute one iteration.)

C4 (Initialize within each iteration). Set $n \leftarrow 1$, $t \leftarrow coeff\, y^a(1-y)^b$, $s \leftarrow at/y/(1-y)$, $q \leftarrow q_0$, $v \leftarrow q$, $CDF \leftarrow vt$, $PDF \leftarrow qs$.

C5 (Compare $a+n$, $(a+b+n)y$). If $a+n > (a+b+n)y$, go to step C7.

C6 (Update the term and then accumulate for both of $F(y; m, N, \rho^2)$ and $f(y; m, N, \rho^2)$. Then increase n by 1). Set $q \leftarrow q(a+b+n-1)(\rho^2)/n$, $v \leftarrow v+q$, $s \leftarrow t(a+b+n-1)/(1-y)$, $t \leftarrow ty(a+b+n-1)/(a+n)$, $CDF \leftarrow CDF + vt$, $PDF \leftarrow PDF + qs$, $n \leftarrow n+1$, and return to step C5.

C7 (Find the corresponding error bounds). Set $bndcdf \leftarrow ty(a+b+n-1)/((a+n)-(a+b+n)y)$, $bndpdf \leftarrow t(a+b+n-1)(1-v)/(1-y)$.

C8 (Check for convergence). If $bndcdf \leq EPS$ and $bndpdf \leq EPS$, go to step C10.

C9 (Update the terms and then accumulate for $F(y; m, N, \rho^2)$ and/or $f(y; m, N, \rho^2)$. Then increase n by 1). Set $q \leftarrow q(a+b+n-1)(\rho^2)/n$, $v \leftarrow v+q$. If $bndcdf \leq EPS$, set $s \leftarrow sy(a+b+n-1)/(a+n-1)$, $PDF \leftarrow PDF + qs$, $n \leftarrow n+1$, $bndpdf \leftarrow sy(a+b+n-1)(1-v)/(a+n-1)$, and return to step C8; otherwise if $bndpdf \leq EPS$, set $t \leftarrow ty(a+b+n-1)/(a+n)$, $CDF \leftarrow CDF + vt$, $n \leftarrow n+1$, $bndcdf \leftarrow ty(a+b+n-1)/((a+n)-(a+b+n)y)$, and return to step C8; otherwise set $s \leftarrow t(a+b+n-1)/(1-y)$, $t \leftarrow ty(a+b+n-1)/(a+n)$, $CDF \leftarrow CDF + vt$, $PDF \leftarrow PDF + qs$, $n \leftarrow n+1$, and return to step C7.

C10 (Find new y and check for convergence of Newton's process). Set $diff \leftarrow (CDF - p)/PDF$. If $y - diff \leq 0$, set $y \leftarrow y/2$; otherwise if $y - diff \geq 1$, set $y \leftarrow (y+1)/2$; otherwise set $y \leftarrow y - diff$. If $|diff|/y \leq DELTA$, terminate the algorithm. (y is the answer.)

C11 (Output error message). Terminate the algorithm with the message "No convergence after $N_{\max}$ iterations".

FORTRAN codes based on the above algorithms are available upon request.

References

Abramowitz, M. and I.A. Stegun, *Handbook of mathematical functions* (Dover, New York, 1965).

Anderson, T.W., *An introduction to multivariate analysis*, 2nd. Edn. (Wiley, New York, 1984).

Ding, C.G., On the computation of the noncentral beta distribution, *Comput. Statist. Data Anal.*, **18** (1994) 449–455.

Ding, C.G. and R.E. Bargmann, Algorithm AS 260: evaluation of the distribution of the square of the sample multiple-correlation coefficient, *Appl. Statist.*, **40** (1991a) 195–198.

Ding, C.G. and R.E. Bargmann, Algorithm AS 261: quantiles of the distribution of the square of the sample multiple-correlation coefficient, *Appl. Statist.*, **40** (1991b) 199–202.

Kennedy, W.J. and J.E. Gentle, *Statistical computing* (Marcel Dekker, New York, 1980).