A Fortran 90 Program
for
Evaluation of Multivariate Normal and Multivariate t Integrals
Over Convex Regions

Paul N. Somerville
Department of Statistics
University of Central Florida
Orlando, FL USA 32816

0. Introduction

Let **X**' = (X_1,X_2, ... ,X_k) have the multivariate normal distribution f(**X**) = MVN($\boldsymbol{\mu}$, $\Sigma\sigma^2$) where Σ is a known positive definite matrix, and σ^2 is a constant. There are many problems in statistics which require the evaluation of f(x) over some convex region A. That is

$$P = \int_A f(\mathbf{X})\, d\mathbf{X}.$$

If σ^2 is known, then without loss of generality, set $\boldsymbol{\mu} = \mathbf{0}$, σ =1 and let Σ be the correlation matrix. For the case where the region A is rectangular, the problem has been addressed by many authors. They include Gupta (1963), Milton (1972), Schervish (1984), Deak (1986), Wang and Kennedy (1990,1992), Olson and Weissfeld (1991), Drezner (1992) and Genz (1992,1993). However, regions of integration for many statistical applications, for example multiple comparisons, are not rectangular.

If σ^2 is estimated by s^2 with ν degrees of freedom such that $\nu\, s^2 / \sigma^2$ is a chi-square variate, then without loss of generality we may assume f(**X**) has the central multivariate-t distribution with correlation matrix Σ. For the case where the correlations have the product form $\rho_{ij} = \lambda_i \lambda_j$ $(i \neq j)$, Dunnett (1989) developed an algorithm to evaluate integrals of the multivariate t over rectangular regions.

The Fortran 90 program MVI3.FOR (which combines and extends two previous programs, MVI and MVIB) can be used to evaluate multivariate normal or multivariate-t integrals over any region which is bounded by hyperplanes (and thus convex). Three decimal accuracies may be obtained in seconds for 20 dimensional integrals using a 486 processor.

The original programs MVI and MVIB could be used only if the convex region contained the origin. The present program MVI3 has no such limitation. However, if the region does not contain the origin, MVI3 uses “crude Monte Carlo” for the evaluation. In addition, the user of the program may elect to do the evaluation using “crude Monte Carlo” even when the convex region contains the origin.

1. Methodology

The following is an overview of the methodology involved.

With no loss of generality, we can assume that the mean is at the origin of the coordinate system. A coordinate transformation (Cholesky) is made so that in the new coordinate system we have k uncorrelated normal or spherically symmetric t random variables with unit variances. The region will be bounded by the transformed hyperplanes.

We discuss first the case where the origin is inside the convex region. Consider a randomly selected point in the k dimensional space. The square of the distance to the point is the sum of the squares of the individual coordinates of the point. Thus if σ^2 is known, it is distributed as χ^2 with k degrees of freedom. If σ^2 is not known, the distance squared divided by k has the F distribution with k degrees of freedom for the numerator and degrees of freedom for the denominator the same as those for the estimate of σ^2.

Select at random a direction in the k dimensional space and obtain the distance from the origin to the boundary. An unbiased estimate of the value of the integral is the probability of obtaining a distance less than or equal to the distance to the boundary.

Using the program (MVI), a prespecified number of random directions, say mocar, are obtained. The average of the corresponding probabilities estimates the value of the integral. To obtain a standard error, repeat the process of obtaining an estimate using mocar random directions a specified number of times (say irep). The final integral estimate is the average of the irep estimates, and the standard error of this estimate is obtained from the standard error of the irep individual estimates.

The program (MVIB) uses a somewhat more efficient procedure (called binning). Instead of calculating a probability corresponding to each individual random direction, the distances are "binned" and an empirical frequency step function of the distances is obtained. Gauss Legendre quadrature is then used to obtain a single estimate of the value of the integral for the mocar random directions. The boundaries of the bins are chosen so as to optimize the quadrature.

The binning procedure has been shown to be especially useful, Somerville (1997), in the calculation of percentage points for a statistic, using an iterative procedure. Using the "binning" procedure the random directions need only be generated for the first iteration. The empirical distribution function generated by the first iteration is used for subsequent iterations and the time for subsequent iterations is negligible.

The functions of the two programs MVI and MVIB have been combined in the program MVI3. In addition, the program has been extended to include the case where the origin is not in the region. For this case, "crude Monte Carlo" is used. Random points in the k space are selected and a determination is made as to whether the point is in or out of the region. To increase the efficiency of the Monte Carlo procedure, the boundaries are sorted and ordered so that the "signed distances" from the origin to the boundaries are non-decreasing. The efficiency of the Monte Carlo procedure is greatest for small integral values since the variance of the estimate decreases as the value ot the integral decreases toward zero

The user may elect to use "crude Monte Carlo", regardless of whether or not the region contains the origin.

A comprehensive description of MVI and MVIB is given in Somerville and Wang (1994) and Somerville (1998). The latter paper also contains a comparison of the efficiencies of the methods. A summary is given in the APPENDIX.

2. User instructions and examples for MVI3

The program requires that two files exist before the program is run. QCALC.in contains the input requirements and QCALC.out is used for the program output. QCALC.in can contain the instructions for several integral evaluations, in which case the evaluations are done sequentially, with QCALC.out containing the results for each of the evaluations.

The instructions for a given evaluation consists of three parts for a total of k + mm +1 lines in QCALC.in.

Part 1 One line consisting of 6 (integer) items (in order): k, iseed, ndenom, mocar, irep, mm.

k dimension of the integral
iseed integer in range 1 to 2^{31} - 1 for a 32 bit machine
ndenom degrees of freedom for variance estimate (use -1 if variance known)
mocar # of random directions used for each individual estimate
irep # of individual estimates
mm # of hyperplanes boundaries for the convex region

<u>Part 2</u> k lines

one line is used for each of the k rows of the lower triangular portion of **either** the variance covariance **or** the correlation matrix.

<u>Part 3</u> mm rows of k + 1 constants

one row is used for each hyperplane. Each row contains the coefficients of the expression **l' x** $\le$ d, where **l** and **x** are vectors.

We give two examples to explain the mm rows of constants. Suppose k=3, and the ranges for x_1, x_2 and x_3 are (-1, 2), (-2.1, 1.4) and (-.5, ∞) respectively. Then the (mm = 5) boundaries are:

$x_1 \ge -1$ $\quad x_2 \ge -2.1$ $\quad x_3 \ge -.5$
$x_1 \le 2$ $\quad x_2 \le 1.4$

However, the program requires $\le$ for each boundary, and thus we use:

$-x_1 \le 1$ $\quad -x_2 \le 2.1$ $\quad -x_3 \le .5$
$x_1 \le 2$ $\quad x_2 \le 1.4$

The input for **<u>Part 3</u>** becomes:

```
-1  0  0   1
 1  0  0   2
 0 -1  0   2.1
 0  1  0   1.4
 0  0 -1   .5
```

For our second example, suppose there are four boundaries (mm = 4) for the region of integration:

$x_1 + x_2 \le 3$
$x_2 + x_3 \ge -2$
$x_1 + x_2 + x_3 \le 2.5$
$x_1 - x_2 + x_3 \ge -2.4$

Now $-x_2 - x_3 \le 2$ is equivalent to the second boundary, and $-x_1 + x_2 - x_3 \le 2.4$ is equivalent to the fourth boundary.The input for **<u>Part 3</u>** for the second example becomes:

```
 1  1  0   3
 0 -1 -1   2
 1  1  1   2.5
-1  1 -1   2.4
```

Returning to the first example, if we wished 10 different estimates, each using 1000 random directions, σ^2 was estimated with 30 degrees of freedom, and we used the seed 457 for the random number generator, then the complete input for QCALC.in would be:

Part 1 3 457 30 1000 10 5
Part 2 1
0 1
0 0 1
Part 3

```
-1  0  0   1
 1  0  0   2
 0 -1  0   2.1
 0  1  0   1.4
 0  0 -1    .5
```

To use the binning procedure (MVIB), we need only to make mm negative. Then **Part 1** is

3 457 -1 1000 10 -5

To evaluate the integral using “crude Monte Carlo” methods, we make k negative. **Either** of the following first lines (**Part 1**) could be used:

```
        -3 457 -1 1000 10  5
        -3 457 -1 1000 10 -5
```

Several integrals may be evaluated in a single run. The k+mm+1 lines are entered sequentially in QCALC.in. The program expects a line of 4 or more -1’s after the final evaluation request i.e. -1 -1 -1 -1 -1 -1.

The output (in QCALC.out) for the above example would be:

```
date and time 19981008102335.520
 elapsed time in seconds is      1.000
 no of pops is           3 seed is        457
  df for var est is         30
 no of ran dir. is       1000
 value of integral is calculated         10 times
 mean value is    5.003167E-01
 standard error of mean is    5.807164E-04
 --------------------
```

3. Additional Examples

Two possible replacements for the first line might be :

```
   3  457  30  10000   1  5
      or
   3  457  30  100  100   5.
```

The first replacement would use 10000 x 1 = 10000 random directions to produce a single estimate of the 3 dimensional integral. There would be no estimate of the accuracy of the calculation for the integral. Using the second replacement, we would use 100 x 100 = 10000 random direction for the estimate of the value of the integral. In practice, using the same seed, the estimate of the value of the integral would be nearly identical for all 3 cases. The program input (in QCALC.in) for the two replacement inputs follows:

```
3 457 30 10000 1 5
1
0 1
0 0 1
-1  0  0  1
 1  0  0  2
 0 -1  0  2.1
 0  1  0  1.4
 0  0  -1 .5

3 457 30 100 100 5
1
0 1
0 0 1
-1  0  0  1
 1  0  0  2
 0 -1  0  2.1
 0  1  0  1.4
 0  0  -1 .5
```

The resulting output was (in QCALC.out)

```
date and time 19981008102336.830
```

```
 elapsed time in seconds is      2.000
 no of pops is           3 seed is        457
  df for var est is         30
 no of ran dir. is      10000
 value of integral is calculated          1 times
 mean value is    5.003167E-01
 --------------------
 date and time 19981008102338.100
 elapsed time in seconds is       .000
 no of pops is           3 seed is        457
  df for var est is         30
 no of ran dir. is        100
 value of integral is calculated        100 times
 mean value is    5.003167E-01
 standard error of mean is    9.492344E-04
 --------------------
```

We give three additional examples of input and output.
Input

```
4 457 -1 100 100 5
 1
.5  1
.5 .5  1
.5 .5 .5  1
 2 -1  0  0  1
 1  0 -1  0  1
 0  0 -1  1  1
-1 -1  2  0  1
-1 -1 -4  0  1

3 457 30 10000 1 -6
1
0 1
0 0 1
 .2865  -.2865  0     1
 .2865  0      -.2865 1
 0      .2865  -.2865 1
-.2865  .2865   0     1
-.2865  0       .2865 1
 0     -.2865   .2865 1

-3 457 30 10000 1 6
1
0 1
0 0 1
 .2865  -.2865  0     1
 .2865  0      -.2865 1
 0      .2865  -.2865 1
-.2865  .2865   0     1
-.2865  0       .2865 1
 0     -.2865   .2865 1

-1 -1 -1 -1 -1 -1
```

Corresponding output:

```
date and time 19981008102338.700
 elapsed time in seconds is      1.000
 no of pops is           4 seed is        457
  variance assumed known
 no of ran dir. is        100
 value of integral is calculated        100 times
 mean value is    1.811718E-01
 standard error of mean is    1.828163E-03
 --------------------
 date and time 19981008102339.200
 Binning method used
 elapsed time in seconds is       .000
 no of pops is           3 seed is        457
  df for var est is         30
 no of ran dir. is      10000
 value of integral is calculated          1 times
 mean value is    9.499553E-01
 --------------------
 date and time 19981008102339.420
 Crude Monte Carlo is used
 elapsed time in seconds is      1.000
 no of pops is          3 seed is        457
 df for var est is          30
 no of random directions is       10000
 points =    20000.000000
 estimate of integral is    9.491000E-01
 standard error is    1.554175E-03
 --------------------
```

4. Comments and comparison of the methods

A large number of experiments were conducted to evaluate the accuracy and validity of the procedures, Somerville (1998). Defining efficiency to mean the ratio of running times to achieve the same standard error of the estimate, MVIB was found to be more efficient than MVI by ratios of 1.26 to 2.25, depending on the dimension of the integral and the number of boundaries. Except for small probabilities, and possibly some “extreme” boundaries, “crude Monte Carlo” is not efficient.

A large number of runs were made to determine the proper number of bins for the binning procedure (MVIB). Except for cases of “extreme” regions, nq = 25 was found to be more than a sufficient number. By “extreme” regions we mean regions departing drastically from a “spheroid with center at the origin”, and especially those with a boundary close to the origin. The proper number of bins required to obtain accuracies consistent with the obtained standard error of the estimate was found to be unrelated to the degrees of freedom for the estimate of the variance, but highly dependent on the minimum distance r* of the boundary from the origin. The following empirical formula was developed relating the value of the number of bins to the minimum distance:

nq* = -14.9 - 39.4*log10(r*) + 21.8*(log10(r*))**2.

Using the “binning” procedure, the program uses nq = 25 whenever nq* ≤ 25. This correspond to a distance of .20 (approximately). For nq* > 25, the program uses nq = nq* +10, unless the distance r* is less than .01, in which case the binning procedure is abandoned and the calculation is done using the MVI methodology.

A comment is in order where one of the boundaries goes through the origin (e.g. **l’x** ≤ 0). The program automatically makes the evaluation using Monte Carlo whenever a boundary passes through the origin. An alternate (and recommended) solution would be to change the boundary from **l’x** ≤ 0 to **l’x** ≤ .000,000,1. Since the integral of the standard normal variable from 0 to

.000,000,1 is 4 x 10^{-8}, the additional error in the estimate of the integral value can increase by no more than that amount.

To generate random normal deviates, MVI3 uses an efficient generator recently developed by Marsaglia (1998).

A question which arises is how many random directions should be chosen. Using 10,000 random directions should always result in an estimate whose standard error is less than .005 and may be sufficient for many applications. The standard errors for integral values near 0, and especially for integral values near 1 are much smaller. Three decimal accuracies are obtained in seconds for 20 dimensional integrals on a 486 processor.

5. References

Deak, I. (1986). Computing probabilities of rectangles in case of multinormal distribution. *J. Statist.Comp. and Simul.* **26**, 101-114.

Drezner, Z. (1992). Computation of the multivariate normal integral. *ACM Transactions on Mathematical Software* **18**, 470-480.

Dunnett, C.W. (1989). Multivariate normal probability integrals with product moment correlation structure. Algorithm AS 251. *Applied Statistics* **38**, 564-579. Correction note. *Applied Statistics* **42**, 709.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *J. Graph. and Comp. Statist.* **1**, 141-149.

Genz, A. (1993). Comparison of methods for the computation of multivariate normal probabilities. *Proceedings of the 25th Symposium on the Interface*, San Diego, April 1993.

Gupta, S.S. (1963). Probability integrals of multivariate normal and multivariate t. *Ann. Math. Statist.*, **34**, 792-828.

Marsaglia, G. (1998). Personal communication.

Milton, R.C. (1972). Computer evaluation of the multivariate normal integral *Technometrics*, **14**, 881-889.

Olson, J.M. and Weissfeld, L.A. (1991). Approximation of certain multivariate integrals. *Statistics and Probability Letters*, **11**, 309-317.

Press,William H. et al (1986, 1992) Numerical methods in Fortran, the art of scientific computing, *Cambridge University Press*.

Schervish, M. (1984). Multivariate normal probabilities with error bound. *Applied Statistics*, **33**, 81-87.

Somerville, P. N. and Wang, M.C. (1994). Computation of multivariate normal probabilities over convex regions. *Proceedings of the 26th Symposium on the Interface*, *Computing Science and Statistics*, June 1994, 229-231.

Somerville, Paul N. (1997). Multiple testing and simultaneous confidence intervals: calculation of constants. *Computational Statistics and Data Analysis* **25**, 217-223.

Somerville, Paul N. (1998) Numerical computation of multivariate normal and multivariate-t probabilities over convex regions. *Journal of Computational and Graphical StatisticsI*, Vol.**7**, No. **4**, 529-544.

Wang. M.C. and Kennedy, W.J. (1992). A numerical method for accurately approximating multivariate normal probabilities. *Comp. Statist. and Data Anal.* **13**, 197-210

APPENDIX

Let $\mathbf{X}' = (X_1, X_2, \ldots, X_k)$ have the multivariate normal distribution $f(\mathbf{X}) = MVN(\boldsymbol{\mu}, \Sigma\, \sigma^2)$ where Σ is known and σ^2 is a constant. We may assume without loss of generality that $\mathbf{X}$ is $MVN(\mathbf{0}, \Sigma\, \sigma^2)$. We wish to evaluate

$$P = \int_A f(\mathbf{X})\, d\mathbf{X}$$

where A is the convex region, containing the origin, and bounded by mm ($\geq$ 1) hyperplanes described by

$$L\mathbf{x} \leq \mathbf{d}$$

where $\mathbf{L}' = (\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_k)$. The j^th^ hyperplane is given by $\mathbf{l}_j'\, \mathbf{x} = d_j$.
Let $\Sigma = T\,T'$ (Cholesky decomposition) and set

$$\mathbf{X} \;= T\, \mathbf{W}.$$

The random variables $W_1, W_2, \ldots, W_k$ are independent standard normal or spherically symmetric t variables. With the transformation, the region becomes

$$\mathbf{Gw} \leq \mathbf{d}$$

where

$$\mathbf{G} \;= \mathbf{L}\, T.$$

Setting $\mathbf{G}' = (\mathbf{g}_1, \mathbf{g}_2, \ldots .\mathbf{g}_k)$, the j^th^ hyperplane becomes

$$\mathbf{g}_j'\, \mathbf{w} = d_j.$$

Consider a randomly selected point in the k dimensional space. The square of the distance to the point is the sum of the squares of the individual cordinates (random N(0,1) variables) of the point on the boundary. Thus if σ^2 is known, it is distributed as χ^2 with k degrees of freedom. If σ^2 is not known, the distance squared divided by k has the F distribution with k degrees of freedom for the numerator and degrees of freedom for the denominator the same as those for the estimate of σ^2.

Select at random a direction in the k dimensional space and obtain the distance from the origin to the boundary. An unbiased estimate of the value of the integral is the probability of obtaining a distance less than or equal to the distance to the boundary. Then, for a random direction, if σ^2 is unknown, an unbiased estimate of the integral P is

$$\text{Prob}\,[\, F \leq r^2/k].$$

If σ^2 is known, the unbiased estimate is

$$\text{Prob}\,[\, \chi^2 \leq r^2].$$

To implement the procedure, successive random directions are chosen and corresponding estimates obtained. The value of the integral is the arithmetic mean of the individual estimates.

Binning procedure (MVIB)

Let r* be the minimum distance from the origin to the boundary of A. Divide A into two regions, the portion inside the hypersphere of radius r* and centered at the origin, and the region outside (say A_2). For σ^2 unknown, the probability content of the hypersphere is

$$P_1 = \text{Prob}\,[\, F \leq r^{*2}/k].$$

For σ^2 known the probability is

$$P_1 = \text{Prob}\,[\, \chi^2 \leq r^{*2}].$$

As before let r be the distance from the origin to a randomly selected point in the k-space. The relative frequency function of r, h(r) may be readily derived from the chi-square or F distributions depending on whether σ^2 is known or must be estimated. Let R be the distance from the origin to the boundary of the region in a random direction. Let G(R) be the cumulative distribution function of R. Then the probability content of the region is

$$P_2 = \int [1 - G(r)]\, h(r)\, dr$$

where the integration is from r* to infinity. The cumulative distribution function G(R) can be estimated using Monte Carlo.

It will be convenient to use reciprocal distance instead of distance. Putting v = 1/r, we obtain

$$P_2 \quad = \int E(v)\, g(v)\, dv$$

where the integration is from 0 to 1/r*. The relative frequency function g(v) may be derived from h(r). E(v) = 1 - G(r) is the cumulative distribution function of the reciprocal distance from the origin to the boundary of the region.

Let a_i be the i^{th} abscissa and wt_i be the corresponding weight for the q-point Gauss-Legendre quadrature. For each random direction **c**, the corresponding distance d is calculated and by means of a binning process, estimates are obtained for $E(a_i)$, i = 1 to q. The estimate of P_2 is given by

$$P_2 \quad = \Sigma\, E(a_i)\, g(a_i)\, wt_i$$

where the summation is from i = 1 to q.