

# The Generalized $F$ Distribution

Donald E. Ramirez  
Department of Mathematics  
University of Virginia  
Charlottesville, VA 22903-3199 USA  
der@virginia.edu  
<http://www.math.virginia.edu/~der/home.html>

*Key Words:* Generalized  $F$  distribution, Cook's  $D_I$  statistic, outliers, misspecified Hotelling's  $T$  test, GENF.

## 1 Introduction

This paper discusses an algorithm for computing the cumulative distribution function *cdf* for the generalized  $F$  distribution and the companion Fortran77 code GENF. Examples of such distributions are the Cook's  $D_I$  statistics and the Hotelling's  $T$  test when the covariance is misspecified.

### 1.1 Basic Results

Cook's (1977)  $D_I$  statistics are used widely for assessing influence of design points in regression diagnostics. These statistics typically contain a leverage component and a standardized residual component. Subsets having large  $D_I$  are said to be influential, reflecting high leverage for these points or that  $I$  contains some outliers from the data. Consider the linear model

$$\mathbf{Y}_0 = \mathbf{X}_0\beta + \varepsilon_0 \tag{1}$$

where  $\mathbf{Y}_0$  is a  $(N \times 1)$  vector of observations,  $\mathbf{X}_0$  is a  $(N \times k)$  full rank matrix of known constants,  $\beta$  is a  $(k \times 1)$  vector of unknown parameters, and  $\varepsilon_0$  is a  $(N \times 1)$  vector of randomly distributed Gaussian errors with  $E(\varepsilon_0) = \mathbf{0}$  and  $Var(\varepsilon_0) = \sigma^2\mathbf{I}_N$ . The least squares estimate of  $\beta$  is  $\hat{\beta} = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{Y}_0$ . The basic idea in *influence analysis* as introduced by Cook (1977) concerns the stability of a linear regression model under small perturbations. For example, if some cases are deleted, then what changes occur in estimates for the parameter vector  $\beta$ ? Cook's  $D_I$  statistics are based on a Mahalanobis distance between  $\hat{\beta}$  (using all the cases) and  $\hat{\beta}_I$  (using all cases except those in the subset  $I$ ), as given by

$$D_I(\hat{\beta}, \mathbf{M}, c\hat{\sigma}^2) = (\hat{\beta}_I - \hat{\beta})' \mathbf{M}(\hat{\beta}_I - \hat{\beta}) / (c\hat{\sigma}^2), \tag{2}$$

with  $\mathbf{M}$  a  $(k \times k)$  nonnegative definite matrix,  $\hat{\sigma}^2$  is an unbiased estimate of the variance, and  $c$  a user defined constant. We use the estimator  $s_I^2$ , the sample variance estimator with the cases in  $I$  omitted, and we use  $c = r$ . We will discuss the cases with  $\mathbf{M} = \mathbf{X}'_0 \mathbf{X}_0$  and  $\mathbf{X}' \mathbf{X}$ , where  $\mathbf{X}$  denotes the remaining rows of  $\mathbf{X}_0$ , and  $\mathbf{A}^+$  denotes the Moore-Penrose generalized inverse of  $\mathbf{A}$ . Let

$$Q_I(\hat{\beta}, \mathbf{M}) = (\hat{\beta}_I - \hat{\beta})' \mathbf{M} (\hat{\beta}_I - \hat{\beta}) \quad (3)$$

denote the quadratic form in the numerator of Equation 2. We have chosen  $s_I^2$  as the estimator for  $\sigma^2$  since this estimator and  $Q_I$  of Equation 3 are independent.

To use  $D_I$  diagnostically, Cook (1977) and Weisberg (1980, p. 108) suggested using the 50th percentile from the central  $F$  distribution with degrees of freedom  $(k, N - k)$  as a benchmark for identifying influential subsets. Since  $D_I$  is not distributed as a central  $F(k, N - k)$  distribution (since  $\hat{\beta}_I$  and  $\hat{\beta}$  are correlated), they recommended the 50th percentile as a rule-of-thumb for determining influential observations. Later Jensen and Ramirez (1998a) derived the *cdf*'s of  $D_I$  as generalized  $F$  distributions. Using the algorithm for computing the generalized  $F$  distribution discussed in this paper, we are able to numerically compute the *cdf* of Cook's  $D_I$  statistics, and, in particular, to compute the  $p$ -values for  $D_I$ . This approach provides a statistical procedure for identifying influential observations based on  $p$ -values.

## 1.2 Notation

To fix the notation, let  $I$  be a subset of  $\{1, \dots, N\}$ , say  $I = \{i_1, \dots, i_r\}$ . Let  $\mathbf{X}_0$  be partitioned as  $\mathbf{X}'_0 = [\mathbf{X}', \mathbf{Z}']$ , with  $\mathbf{X}$  containing the rows determined by  $I$ , and  $\mathbf{Z}$  the remaining rows. We assume that the matrices  $\mathbf{X}_0$ ,  $\mathbf{X}$ , and  $\mathbf{Z}$  all of full rank, of orders  $(N \times k)$ ,  $(n \times k)$ , and  $(r \times k)$ , respectively such that  $k < n < N$ , and  $n + r = N$ , with  $r < k$  for notational convenience. Partition  $\mathbf{Y}'_0 = [\mathbf{Y}'_1, \mathbf{Y}'_2]$ , and  $\varepsilon'_0 = [\varepsilon'_1, \varepsilon'_2]$ . Thus Equation 1 has been transformed into

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Z} \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}. \quad (4)$$

Jensen and Ramirez (1998a) adapted the theory of singular decompositions to transform the linear model Equation 4 into canonical form. They constructed orthogonal matrices  $\mathbf{Q}_1 \in \mathcal{O}(n)$  and  $\mathbf{Q}_2 \in \mathcal{O}(r)$  and a nonsingular matrix  $\mathbf{G}$  such that  $\mathbf{Q}_1 \mathbf{X} \mathbf{G} = [\mathbf{I}'_k \mathbf{0}]'$  and  $\mathbf{Q}_2 \mathbf{Z} \mathbf{G} = [\mathbf{D}_\gamma \mathbf{0}]$ , where  $\mathbf{D}_\gamma$  is the diagonal matrix whose elements  $\{\gamma_1 \geq \dots \geq \gamma_r > 0\}$  comprise the square roots of the ordered eigenvalues of  $\mathbf{Z}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{Z}'$ . The ordered eigenvalues of  $\mathbf{Z}(\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{Z}'$  are denoted  $\{\lambda_1 \geq \dots \geq \lambda_r > 0\}$  with  $\{\lambda_i = \gamma_i^2 / (1 + \gamma_i^2); 1 \leq i \leq r\}$  the canonical leverages.

The matrices  $\mathbf{Q}_1$ ,  $\mathbf{Q}_2$ , and  $\mathbf{G}$  are used to transform the original model  $\mathbf{Y}_0 =$

$\mathbf{X}_0\beta + \varepsilon_0$  into canonical form

$$\begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \\ \mathbf{U}_3 \\ \mathbf{U}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_s \\ \mathbf{0} & \mathbf{0} \\ \mathbf{D}_\gamma & \mathbf{0} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} \quad (5)$$

where  $\mathbf{Q}_1\mathbf{Y}_1 = [\mathbf{U}'_1, \mathbf{U}'_2, \mathbf{U}'_3]'$ ,  $\mathbf{Q}_1\varepsilon_1 = [\eta'_1, \eta'_2, \eta'_3]'$ , with  $(\mathbf{U}_1, \eta_1) \in \mathbb{R}^r$ ,  $(\mathbf{U}_2, \eta_2) \in \mathbb{R}^s$ ,  $(\mathbf{U}_3, \eta_3) \in \mathbb{R}^t$ , such that  $r + s = k$  and  $t = n - k$ . Further  $\mathbf{Q}_2\mathbf{Y}_2 = \mathbf{U}_4$  and  $\mathbf{Q}_2\varepsilon_2 = \eta_4$  with  $(\mathbf{U}_4, \eta_4) \in \mathbb{R}^r$  and  $[\theta'_1, \theta'_2]' = [(\mathbf{G}^{-1}\beta_1)', (\mathbf{G}^{-1}\beta_2)']'$ . For the reduced data in canonical form

$$\begin{bmatrix} \hat{\theta}_{I1} \\ \hat{\theta}_{I2} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix}, \quad (6)$$

and for the full data

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} (\mathbf{I}_r + \mathbf{D}_\gamma^2)^{-1}(\mathbf{U}_1 + \mathbf{D}_\gamma\mathbf{U}_4) \\ \mathbf{U}_2 \end{bmatrix} \quad (7)$$

with  $\text{cov}(\hat{\theta}_{I1} - \hat{\theta}_1) = \sigma^2 (\mathbf{I}_r - (\mathbf{I}_r + \mathbf{D}_\gamma^2)^{-1}) = \sigma^2 \mathbf{D}_\gamma (\mathbf{I}_r + \mathbf{D}_\gamma^2)^{-1} \mathbf{D}'_\gamma$ .

We now define the *generalized F distribution*. Suppose that the elements of  $\mathbf{U} = [U_1, \dots, U_r]'$  ( $r > 1$ ) are independent  $\{N_1(0, 1); 1 \leq i \leq r\}$  random variables; let  $\{\alpha_1, \dots, \alpha_r\}$  be nonincreasing positive weights; and identify  $T = \alpha_1 U_1^2 + \dots + \alpha_r U_r^2$ . If  $\mathcal{L}(V) = \chi^2(\nu)$  independently of  $\mathbf{U}$ , then the *cdf* of  $W = (T/r)/(V/\nu)$  is denoted by  $F_r(w; \alpha_1, \dots, \alpha_r; \nu)$ . If all of the  $\alpha_i$  ( $1 \leq i \leq r$ ) are equal to say  $\alpha$ , then the *cdf* of  $W$  is denoted by  $F_r(w; \alpha; \nu)$ , the scaled central  $F$  distribution with degrees of freedom  $(r, \nu)$ . If  $\{\mathcal{L}(U_i) = N_1(\omega_i, 1); 1 \leq i \leq r\}$ , then the *cdf* of  $W$  is denoted by  $F_r(w; \alpha_1, \dots, \alpha_r; \omega_1, \dots, \omega_r; \nu)$ , the *noncentral generalized F distribution*.

Jensen and Ramirez (1991) showed that the *cdf* for  $W_0 = T/V$ , equivalently for  $W = (T/r)/(V/\nu)$ , is a weighted series of  $F$  distributions, and they computed the stochastic bounds

$$F_r(w; \alpha_1; \nu) \leq F_r(w; \alpha_1, \dots, \alpha_r; \nu) \leq F_r(w; \alpha^*; \nu), \quad (8)$$

with  $\alpha_1$  the maximum weight,  $\alpha^*$  the geometric mean of the weights  $\{\alpha_1, \dots, \alpha_r\}$ , and  $F_r(w; \alpha; \nu)$  the scaled central  $F$  distribution.

The basic characterization theorems for  $D_I$  are given in Jensen and Ramirez (1998a) and are:

**Theorem 1** Suppose that  $\mathcal{L}(\mathbf{Y}) = N_N(\mathbf{X}_0\beta, \sigma^2\mathbf{I}_N)$ , then

(1) The distribution of  $D_I(\hat{\beta}, \mathbf{X}'_0\mathbf{X}_0, rs_I^2) = D_I(\hat{\theta}_1, \mathbf{I}_r + \mathbf{D}_\gamma^2, rs_I^2)$  is given by  $F_r(w; \gamma_1^2, \dots, \gamma_r^2; N - r - k)$ .

(2) The distribution of  $D_I(\hat{\beta}, \mathbf{X}'\mathbf{X}, rs_I^2) = D_I(\hat{\theta}_1, \mathbf{I}_r, rs_I^2)$  is given by  $F_r(w; \lambda_1, \dots, \lambda_r; N - r - k)$ .

With  $r = 1$ ,  $\mathcal{L}(D_i(\hat{\beta}, \mathbf{X}'_0 \mathbf{X}_0, s_i^2)/\gamma_1^2) = \mathcal{L}(D_i(\hat{\beta}, \mathbf{X}' \mathbf{X}, s_i^2)/\lambda_1) = F(1, N - 1 - k)$  with the two  $p$ -values from Theorem 1 all being equal when  $r = 1$ . Outliers also can be tested using the studentized deleted residuals with  $\mathcal{L}((y_i - \hat{y}_{(i)})/(s_i \sqrt{1 + \mathbf{x}_i(\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i})) = t(N - 1 - k)$  where  $\hat{y}_{(i)}$  denotes the predicted value using  $(\mathbf{Y}_1, \mathbf{X})$ ; or with the externally studentized residuals (*RStudent*) with  $\mathcal{L}((y_i - \hat{y}_i)/(s_i \sqrt{1 - h_{ii}})) = t(N - 1 - k)$  where  $\hat{y}_i$  denotes the predicted value using  $(\mathbf{Y}, \mathbf{X}_0)$  and  $h_{ii}$  is the canonical leverage also denoted as  $\lambda_1$ . In Jensen and Ramirez (1998b) it is shown that the  $p$ -values from these two tests are also equal to the two  $p$ -values from Theorem 1. Thus, in case of single deletion with  $r = 1$ , all of these four standard tests for outliers will have a common  $p$ -value.

This paper concerns the case of joint outliers with  $r > 1$ .

## 2 The Distribution of $T/V$ and $(T/r)/(V/\nu)$

Building on the work of Gurland (1955) and Kotz, Johnson, and Boyd (1967), Ramirez and Jensen (1991) showed how to compute the *cdf* for  $W_0 = T/V$  as a weighted series of  $F$  distributions, and they computed the error bounds for the truncated partial sums. Their results are stated for  $W_0 = T/V$  with  $r = p$  and with  $\mathcal{L}(V) = \chi^2(\nu - p + 1)$ . We give the results for the general case below.

### 2.1 Weighted $F$ Series

For the general case when the  $\alpha_i$  are not all the same value, define recursively the sequences

$$\begin{aligned} c_0 &= \prod_{i=1}^r (\delta/\alpha_i)^{1/2}; \\ d_j &= \frac{1}{2} \sum_{i=1}^r (1 - \delta/\alpha_i)^j, \quad j \geq 1; \\ c_j &= \frac{1}{j} \sum_{l=0}^{j-1} (d_{j-l} c_l), \quad j \geq 1, \end{aligned} \tag{9}$$

where  $\delta$  satisfies  $0 < \delta \leq \alpha_r$ . The program GENF sets  $\delta = 1.0\alpha_r$ . (In GENF, the variable NEAR1 is set equal to 1.0.) This assures that  $0 \leq 1 - \delta/\alpha_i < 1$  ( $1 \leq i \leq r$ ). Set  $\epsilon = \max\{1 - \delta/\alpha_i; 1 \leq i \leq r\}$ , so  $0 < \epsilon < 1$ , since  $\alpha_1 \neq \alpha_r$ . In the case  $\alpha_1 = \dots = \alpha_r = \alpha$ , then  $\epsilon = 0$  and  $F_r(w; \alpha, \dots, \alpha; \nu) = F_r(w; \alpha; \nu)$ , the scaled central  $F$  distribution. In the following sections, we will assume that  $\alpha_1 \neq \alpha_r$ . The program GENF checks for this condition, and if satisfied, then the program computes the  $p$ -value from the scaled central  $F$  distribution.

The *pdf* of  $T/V$  has the representation

$$\begin{aligned} h_{T/V}(w) &= \sum_{j=0}^{\infty} \frac{c_j}{\delta} \frac{\Gamma((\nu+r+2j)/2) (w/\delta)^{(r+2j-2)/2}}{\Gamma((r+2j)/2) \Gamma(\nu/2) (1+w/\delta)^{(\nu+r+2j)/2}} \\ &= \sum_{j=0}^{\infty} \frac{c_j}{\delta} \frac{\nu}{r+2j} f_F\left(\frac{\nu}{r+2j} \frac{w}{\delta}; r+2j, \nu\right) \end{aligned} \quad (10)$$

with  $f_F(w; v_1, v_2)$  the density of the central  $F$  distribution with degrees of freedom  $(v_1, v_2)$ . Equivalently,

$$\begin{aligned} h_{(T/r)/(V/\nu)}(w) &= \sum_{j=0}^{\infty} \frac{r}{\nu} \frac{c_j}{\delta} \frac{\Gamma((\nu+r+2j)/2) (r/w/\delta)^{(r+2j-2)/2}}{\Gamma((r+2j)/2) \Gamma(\nu/2) (1+r/w/\delta)^{(\nu+r+2j)/2}} \\ &= \sum_{j=0}^{\infty} \frac{c_j}{\delta} \frac{r}{r+2j} f_F\left(\frac{r}{r+2j} \frac{w}{\delta}; r+2j, \nu\right). \end{aligned} \quad (11)$$

## 2.2 Bounds

A global truncation error bound  $e_{\tau_0}$  for the  $\tau$ -th partial sum of the *pdf* of  $T/V$  is given by

$$\begin{aligned} e_{\tau_0}(w) &= \sum_{j=\tau+1}^{\infty} \frac{c_j}{\delta} \frac{\nu}{r+2j} f_F\left(\frac{\nu}{r+2j} \frac{w}{\delta}; r+2j, \nu\right) \\ &\leq \frac{\nu}{\delta(r+2(\tau+1))} (1 - (c_0 + \dots + c_{\tau})) = e_{\tau_0} \end{aligned} \quad (12)$$

from the equality  $\sum_{i=0}^{\infty} c_i = 1$ , and since  $|f_F(w; v_1, v_2)| \leq 1$  when  $v_1 \geq 2$  and  $v_2 \geq 1$ . Analytic bounds are given in Ramirez and Jensen (1991) in their Theorems 3.2 and 3.6. They also derived two local bounds (their Theorems 3.3 and 3.7) for the truncation error  $e_{\tau_0}(w)$  for the  $\tau$ -th partial sum of the *pdf* of  $T/V$ . For an integer  $p$ , let  $\langle p \rangle = 2\text{floor}((p+1)/2)$  be the smallest even integer greater than or equal to  $p$ .

Local Bound 1a: For any  $w \geq 0$ , a local error bound for the *pdf* of  $W_0 = T/V$  is

$$e_{\tau_0}(w) \leq \frac{c_0 \epsilon^{\tau+1} \Gamma((2\tau + \nu + r + 2)/2) (w/\delta)^{(r+2\tau)/2}}{\delta \Gamma(r/2) \Gamma(\nu/2) (\tau+1)! (1 + (1-\epsilon)w/\delta)^{(2\tau+\nu+r+2)/2}}. \quad (13)$$

Local Bound 2a: For  $\tau+1 > (\nu+r-2)/(|\log \epsilon|)$  with  $t = (w/\delta)/(1+w/\delta)$ , and with

$$c = \frac{c_0 (w/\delta)^{(r-2)/2}}{\delta \Gamma(r/2) \Gamma(\nu/2) (1+w/\delta)^{(\nu+r)/2}}, \quad (14)$$

then a local error bound for the *pdf* of  $W_0 = T/V$  is

$$e_{\tau 0}(w) \leq \frac{c\Gamma(\langle\nu+r\rangle/2)\epsilon t}{(\epsilon t|\log \epsilon t|)^{\langle\nu+r\rangle/2}} P[V > (2\tau + \langle\nu+r\rangle - 2)|\log \epsilon t|] \quad (15)$$

with  $\mathcal{L}(V) = \chi^2(\langle\nu+r\rangle)$ .

The corresponding inequalities for  $W = (T/r)/(V/\nu)$  are given below. A bound for the global truncation error  $e_\tau$  for the  $\tau$ -th partial sum of the *pdf* of  $W = (T/r)/(V/\nu)$  is given by

$$\begin{aligned} e_\tau(w) &= \sum_{j=\tau+1}^{\infty} \frac{c_j}{\delta} \frac{r}{r+2j} f_F\left(\frac{r}{r+2j} \frac{w}{\delta}; r+2j, \nu\right) \\ &\leq \frac{r}{\delta(r+2(\tau+1))} (1 - (c_0 + \dots + c_\tau)) = e_\tau. \end{aligned} \quad (16)$$

Local Bound 1b: For any  $w \geq 0$ , a local error bound for the *pdf* of  $W = (T/r)/(V/\nu)$  is

$$e_\tau(w) \leq \frac{r}{\nu} \frac{c_0 \epsilon^{\tau+1} \Gamma((2\tau + \nu + r + 2)/2) (\frac{r}{\nu} w/\delta)^{(r+2\tau)/2}}{\delta \Gamma(r/2) \Gamma(\nu/2) (\tau+1)! (1 + (1-\epsilon) \frac{r}{\nu} w/\delta)^{(2\tau+\nu+r+2)/2}}. \quad (17)$$

Local Bound 2b: For  $\tau + 1 > (\nu + r - 2)/(|\log \epsilon t|)$  with  $t = (\frac{r}{\nu} w/\delta)/(1 + \frac{r}{\nu} w/\delta)$ , and with

$$c = \frac{c_0 (\frac{r}{\nu} w/\delta)^{(r-2)/2}}{\delta \Gamma(r/2) \Gamma(\nu/2) (1 + \frac{r}{\nu} w/\delta)^{(\nu+r)/2}}, \quad (18)$$

then a local error bound for the *pdf* of  $W = (T/r)/(V/\nu)$  is

$$e_\tau(w) \leq \frac{r}{\nu} \frac{c\Gamma(\langle\nu+r\rangle/2)\epsilon t}{(\epsilon t|\log \epsilon t|)^{\langle\nu+r\rangle/2}} P[V > (2\tau + \langle\nu+r\rangle - 2)|\log \epsilon t|] \quad (19)$$

with  $\mathcal{L}(V) = \chi^2(\langle\nu+r\rangle)$ .

Figure 1 displays the local truncation error bound using Equation 19 and the global truncation error  $e_\tau = 0.8029 \cdot 10^{-4}$  from Equation 16 for the generalized  $F$  distribution  $F_r(w; \alpha_1, \dots, \alpha_r; \nu)$  with  $r = 3$ ,  $\nu = 9$ , and weights  $\alpha = (2, 2, 1/2)$  using  $\tau = 33$ .

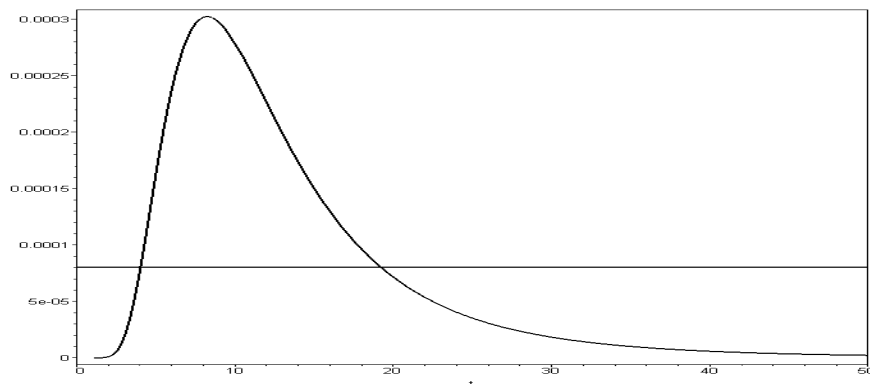


FIGURE 1: LOCAL TRUNCATION ERROR BOUND 2B FOR THE *pdf* OF  $W$

For coding convenience, GENF internally uses Equations 10, 12, 13, and 15 for  $W_0 = T/V$  and transforms the results to the generalized  $F$  distribution  $W = (T/r)/(V/\nu)$ . Thus the input value  $y$  for  $W$  is multiplied by  $r/\nu$  for the internal computation with  $y_0 = \frac{r}{\nu}y$  for calculations with the distribution of  $W_0 = T/V$ . The subroutine GENF will increase  $\tau$  until the global truncation error  $e_{\tau_0}$  from Equation 12 satisfies  $e_{\tau_0} \leq PDFERR$  with  $PDFERR$  a prescribed small value, say  $10^{-4}$ . Usually  $\tau \leq 40$ . If the global truncation error bound is not achieved using  $CSIZE = 3000$  terms, then the error code of  $IER = 8$  is returned. The users would need to increase the value of  $CSIZE$  in the driver program and in the subroutine GENF. The local bounds Equations 12, 13, and 15 are then used to compute the maximum local truncation error  $\max\{e_{\tau_0}(w)\} = ERRDEN0$  over the values  $w$  used by the adapted integration procedure to compute the cumulative distribution function as the integral of the probability density function in Equation 10. To find the density  $h_{(T/r)/(V/\nu)}(y)$  and the maximum local truncation error  $\max\{e_{\tau}(w)\} = ERRDEN$ , the values  $h_{T/V}(y_0)$  and  $ERRDEN0$  are multiplied by  $r/\nu$  respectively. Typically,  $ERRDEN$  provides a tighter bound for the *pdf* error than  $e_{\tau}$  as shown in Figure 1.

### 2.3 Examples

For the Hald (1952, p. 647) data set ( $N = 13$  and  $k = 5$ ) using the test statistic  $D_I(\hat{\beta}, \mathbf{X}'\mathbf{X}, 2s_I^2)$  and the global bounds Equation 8, we can show that the only pair ( $r = 2$ ) of observations (from the 78 possible pairs) which could possibly be influential at the 5% significance level is  $I = \{6, 8\}$  with  $0.0131 < p_I = 0.0218 < 0.0461$  where  $p_I$  is computed with GENF. The inputs are the cardinality  $r = 2$  of  $I$ , the canonical leverages  $\lambda = (0.408676, 0.124019)$  for the weights  $\alpha$ , the degrees of freedom  $\nu = N - r - k = 6$ , and the observed Cook's  $D_I$  statistic  $y = 2.19331$ . The outputs include the  $p$ -value = 0.0218 and the number of terms used  $\tau = 18$ . Additional outputs include the density = 0.0205, the number of function evaluations required in the adaptive integration procedure

EVALS = 63, and the maximum local error bound in the truncated series from Equations 16, 17, and 19  $ERRDEN = 0.9275 \cdot 10^{-4}$ .

For the Longley (1967) data set, Cook (1977) noted that observations 5 and 16 may be influential. To test for the joint influence of  $I = \{5, 16\}$ , we use GENF with  $N = 16$ ,  $k = 7$ , and  $r = 2$ . Using the test statistic  $D_I(\hat{\beta}, \mathbf{X}'\mathbf{X}, 2s_I^2)$ , the inputs are  $r = 2$ , the canonical leverages  $\lambda = (0.690029, 0.614130)$  for the weights,  $\nu = N - r - k = 7$ , and the observed Cook's  $D_I$  statistic  $y = 1.812433$ . The outputs include the  $p$ -value  $p_I = 0.1293$  and the number of terms used  $\tau = 4$ . Additional outputs include the density = 0.1104, the number of function evaluations = 21, and the maximum local truncation error bound  $ERRDEN = 0.1128 \cdot 10^{-5}$ . Using the test statistic  $D_I(\hat{\beta}, \mathbf{X}'\mathbf{X}, 2s_I^2)$  and the global bounds Equation 8, it is easy to compute that the only possible pairs that need to be considered at the 5% significance level are (1)  $I_1 = \{4, 5\}$  with  $0.0382 \leq p_{I_1} = 0.0418 \leq 0.0636$ , (2)  $I_2 = \{4, 15\}$  with  $0.0496 \leq p_{I_2} = 0.0498 \leq 0.0556$ , and (3)  $I_3 = \{10, 16\}$  with  $0.0376 \leq p_{I_3} = 0.0457 \leq 0.0798$  where the  $p$ -values  $p_I$  are computed with GENF.

Our recommendation to the practitioner, who wishes to find joint outliers, is to initially screen for potential joint outliers using Equation 8 with  $D_I(\hat{\beta}, \mathbf{X}'\mathbf{X}, rs_I^2)$  and then to compute, using GENF, the  $p$ -values for  $D_I(\hat{\beta}, \mathbf{X}'\mathbf{X}, rs_I^2)$ . We use  $M = X'X$  since our computer examples show that usually the number of terms  $\tau$  required will be smaller with this choice of  $M$  than with  $M = \mathbf{X}'_0\mathbf{X}_0$ . Note that the ratio  $\lambda_1/\lambda_r = (\gamma_1^2/(1 + \gamma_1^2))/(\gamma_r^2/(1 + \gamma_r^2)) \leq \gamma_1^2/\gamma_r^2$ .

### 3 Misspecified Hotelling's $T$ test

Hotelling's  $T^2$  is used widely in multivariate data analysis, encompassing tests for means, the construction of confidence ellipsoids, the analysis of repeated measurements, and statistical process control. To support a knowledgeable use of  $T^2$ , its properties must be understood when model assumptions fail. Jensen and Ramirez (1991) have studied the misspecification of location and scale in the model for a multivariate experiment under practical circumstances to be described.

To set the notation, let  $N_p(\mu, \Sigma)$  be the Gaussian distribution with mean  $\mu$ , and dispersion  $\Sigma$  and let  $W_p(\nu^*, \Sigma)$  denote the central Wishart distribution having  $\nu^*$  degrees of freedom and scale parameter  $\Sigma$ . Consider the representation  $T^2 = \nu^* \mathbf{Y}'\mathbf{W}^{-1}\mathbf{Y}$  where  $(\mathbf{Y}, \mathbf{W})$  are independent and  $\mathcal{L}(\mathbf{Y}) = N_p(\mu, \Sigma)$  as before, but now  $\mathcal{L}(\mathbf{W}) = W_p(\nu^*, \Omega)$ . Denote the ordered roots of  $\Omega^{-\frac{1}{2}}\Sigma\Omega^{-\frac{1}{2}}$  by  $\{\pi_1 \geq \pi_2 \geq \dots \geq \pi_p > 0\}$ . A principal result for  $T^2$  under misspecified scale is given in Jensen and Ramirez (1991) and is the following.

**Theorem 2** *The statistic  $T^2$  admits a stochastic representation in which  $\mathcal{L}(T^2/\nu^*) = \mathcal{L}(\mathbf{Y}'\mathbf{W}^{-1}\mathbf{Y}) = \mathcal{L}((\pi_1 Z_1^2 + \pi_2 Z_2^2 + \dots + \pi_p Z_p^2)/V)$  such that the following conditions hold:*

- (i)  $\{Z_1, \dots, Z_p, V\}$  are mutually independent,
- (ii)  $\mathcal{L}(Z_i) = N_1(\theta_i, 1)$  where  $\theta = \Sigma^{-\frac{1}{2}}\mu$ , and

(iii)  $\mathcal{L}(V) = \chi^2(\nu^* - p + 1)$ .  
 Equivalently,  $((\nu^* - p + 1)/p)(T^2/\nu^*)$  is the generalized  $F$  distribution  $F_r(w; \pi_1, \dots, \pi_p; \nu^* - p + 1)$ .

### 3.1 MISSPECIFIED SCALE MODEL

For univariate control charts, Student's  $t = \sqrt{N}(\bar{X} - \mu_0)/s_0$  is used to test for shifts of means ( $H_0 : \mu = \mu_0$  against  $H_1 : \mu \neq \mu_0$ ), with  $\{X_1, \dots, X_N\}$  independent random variables distributed as  $\mathcal{L}(X_1) = \dots = \mathcal{L}(X_N) = N_1(\mu_0, \sigma_0^2)$ ; and with  $s_0^2$ , a sample variance random variable, independent of  $\bar{X}$ , computed from  $M$  historical data values, and distributed as  $\mathcal{L}((M - 1)s_0^2/\sigma_0^2) = \chi^2(M - 1)$ . It is reasonable to assume, if the process  $X$  has changed into the process  $Y$ , now with  $\mathcal{L}(Y) = N_1(\mu_1, \sigma_1^2)$ , that *both*  $\mu_0 \neq \mu_1$  and  $\sigma_0^2 \neq \sigma_1^2$ . In this case,  $t$  is no longer a Student  $t(N - 1)$  distribution, but a scaled Student  $t(N - 1)$  distribution,

$$t = \frac{\sqrt{N}(\bar{Y} - \mu_0)}{s_0} \stackrel{H_0}{=} \frac{\sigma_1}{\sigma_0} \frac{(\bar{Y} - \mu_0)/(\sigma_1/\sqrt{N})}{s_0/\sigma_0}. \quad (20)$$

When  $\sigma_0^2 \neq \sigma_1^2$ , Student's  $t$  is misspecified. In particular, when  $\sigma_1^2 > \sigma_0^2$ , if the user uses the tabulated value of  $t(N - 1)$  for the critical value, there will be an increase in Type I error. When  $\sigma_1^2 < \sigma_0^2$ , there will be a corresponding increase of Type II error. Using Equation (20), power curves for  $t$  can be computed for varying values of  $\sigma_1^2/\sigma_0^2$ .

In the following, we extend these results to multivariate control charts. In the multivariate case, the situation is more complicated due to the correlations between the observed variables, and the misspecified Hotelling's  $T^2$  is distributed as a "scaled"  $T^2$  distribution, that is, as a generalized  $F$  distribution.

The conventional model for  $T^2$  is based on a random sample  $\{X_1, \dots, X_N\}$  from  $N_p(\mu, \Sigma)$  using the unbiased sample means and dispersion matrix  $(\bar{\mathbf{X}}, \mathbf{S})$ . We have  $\mathcal{L}(\bar{\mathbf{X}}) = N_p(\mu, \frac{1}{N}\Sigma)$  and  $\mathcal{L}((N - 1)\mathbf{S}) = W_p(N - 1, \Sigma)$ , or  $\mathcal{L}(\frac{N-1}{N}\mathbf{S}) = W_p(N - 1, \frac{1}{N}\Sigma)$ . Thus  $T^2 = (N - 1)(\bar{\mathbf{X}} - \mu)'(\frac{N-1}{N}\mathbf{S})^{-1}(\bar{\mathbf{X}} - \mu) = N(\bar{\mathbf{X}} - \mu)' \mathbf{S}^{-1}(\bar{\mathbf{X}} - \mu)$  and  $\mathcal{L}(((N - p)/p)(T^2/(N - 1))) = F(p, N - p)$ , the central  $F$  distribution when  $N > p$ . See, for example, Timm (1975). If the process dispersion parameters have shifted, then, by Theorem 2,  $T^2$  is misspecified with  $\mathcal{L}((N - 1)\mathbf{S}) = W_p(N - 1, \Omega)$ , and with  $((N - p)/p)(T^2/(N - 1))$  the generalized  $F$  distribution  $F_r(w; \pi_1, \dots, \pi_p; N - p)$ . Here  $r = p$ ,  $\nu = \nu^* - p + 1 = N - p$ , and  $\pi_1 \geq \pi_2 \geq \dots \geq \pi_p > 0$ , the ordered roots of  $\Omega^{-\frac{1}{2}}\Sigma\Omega^{-\frac{1}{2}}$ . Thus power curves for  $T^2$  can be computed for varying values of  $\pi_1 \geq \pi_2 \geq \dots \geq \pi_p > 0$ .

### 3.2 EXAMPLES

An important application of GENF is for computing the power of Hotelling's  $T^2$  test for a multivariate quality control chart. Power analysis for a misspecified mean  $\mu$  is standard. With GENF, the power analysis for a misspecified covariance  $\Omega$  can now be performed. If a process changes, not only will the mean

change but generally the covariance structure will also change. With GENF, the robustness of  $T^2$  under misspecification of scale can be verified by computing the cumulative density of  $T^2$  for varying choices of  $\pi_1 \geq \pi_2 \geq \dots \geq \pi_p > 0$  at the critical value of  $T^2$ . For example, if  $\Omega_\rho$  is a  $3 \times 3$  equicorrelated matrix ( $r = p = 3$ ) with  $\rho = 0.5$ , and if  $\Sigma$  is the identity matrix, then the eigenvalues of  $\Omega_\rho^{-\frac{1}{2}} \Sigma \Omega_\rho^{-\frac{1}{2}}$  are  $\{\pi_1 = (1 - \rho)^{-1}, \pi_2 = (1 - \rho)^{-1}, \pi_3 = (1 + 2\rho)^{-1}\} = \{2, 2, 1/2\}$ . If  $N = 12$  with  $\nu = N - p = 9$ , the nominal 95% critical value of  $((N - p)/p)(T^2/(N - 1))$  is  $F(0.95; p, N - p) = 3.8625$ . However, the exact right-hand tail probability for  $Y = ((N - p)/p)(T^2/(N - 1))$  is not 0.05 but rather  $P[Y = ((N - p)/p)(T^2/(N - 1)) \geq 3.8625] = 0.1231$ , as computed using GENF.

Figure 2 gives the probability distribution function for the generalized  $F$  distribution  $Y = ((N - p)/p)(T^2/(N - 1))$  for the case  $\rho = 0$  and  $\rho = 0.5$  (the misspecified distribution). The misspecified  $T^2$  has the heavier tail.

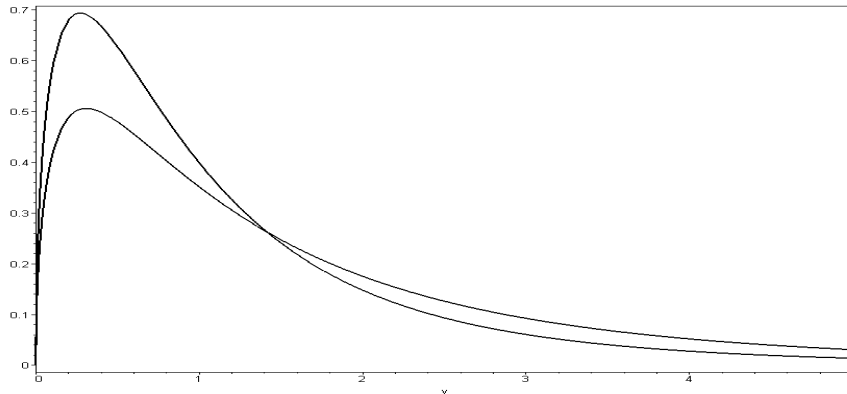


FIGURE 2: GENERALIZED  $F$  PDF FOR  $\rho = 0$  AND  $\rho = 0.5$

In Table 1, we present similar computations for varying  $\rho$ . For each  $\rho$  in the Table 1, and with the corresponding eigenvalues  $\pi_1 \geq \pi_2 \geq \pi_3 > 0$  of  $\Omega_\rho^{-\frac{1}{2}} \Sigma \Omega_\rho^{-\frac{1}{2}}$ , we use GENF to compute the value of  $\tau$  required to satisfy  $ye_\tau \leq 10^{-4}$  and the computed values of  $P[Y = ((N - p)/p)(T^2/(N - 1)) \geq 3.8625]$ . The inputs are  $r = 3$ , the weights  $\pi_1 \geq \pi_2 \geq \pi_3 > 0$ ,  $\nu = N - p = 12 - 3 = 9$ , and  $y = 3.8625$ .

**Table 1. Misspecified Type I Error**

$\rho$	$\tau$	$p = P[Y \geq 3.8625]$
0.0	0	0.0500
0.1	5	0.0527
0.2	8	0.0601
0.3	12	0.0728
0.4	17	0.0926
0.5	24	0.1231
0.6	34	0.1700
0.7	49	0.2458
0.8	77	0.3712
0.9	153	0.5905

The number of terms  $\tau$  is not dependent on the number of terms in the array of positive weights as much as on the ratio of  $\alpha_1/\alpha_r$ , as the following tables show.

For each table the values used were  $\nu = 10$  and  $y = 10$ . The table reports the rank of the array of weights, the values in the array, and the number of terms required in the partial sum expansion using Formula 8. For the last value of  $\tau$ , the value of CSIZE in GENF was changed from 3000 to 4000.

**Table 2. Values of  $\tau$  for varying arrays of weights**

$r$	$\alpha$	$\tau$
5	1 2 3 4 5	29
10	1 2 3 ... 10	80
15	1 2 3 ... 15	148
20	1 2 3 ... 20	233
25	1 2 3 ... 25	334

$r$	$\alpha$	$\tau$
5	10 1 1 1 1	45
5	10 10 10 10 1	75
5	100 1 1 1 1	310
5	100 100 100 100 1	565
5	1000 1 1 1 1	1675
5	1000 1000 1000 1000 1	3469

## 4 THE PROGRAM GENF

The program GENF is a Fortran77 subroutine which requires access to the IMSL subroutines DCHIDF (to evaluate the probability of a chi-squared distribution), DLNGAM (to evaluate the log of the gamma function), DQDAG (to perform an adaptive integration), and DSVRGP (to sort the array of positive weights). Both GENF and the driver program require DFDF (to evaluate the central  $F$  distribution).

The user inputs are  $r$ ,  $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_r > 0$  (the positive weights for the chi-squared distributions which GENF will sort into descending order),  $\nu$ , the value  $y$  for the generalized  $F$  distribution, and the global truncation error criterion (the driver program is set for  $10^{-4}$ ). The program outputs are the left-hand probability of the cumulative distribution function (using the adaptive integration procedure DQDAG), the lower bound  $LB$  and upper bound  $UB$  from Equation 8, the required number of terms  $\tau$  to satisfy the global truncation error  $e_\tau \leq PDFERR$ , the number of function evaluations used, the value of the density function at  $y$ , the maximum of the local truncation error from Equations

16, 17, and 19 over the values used by the integration subroutine, and the error code *IER*.

Since the value of CUMF is computed using a truncated series expansion,  $\text{CUMF} \leq \Pr[Y \leq y]$ . Conversely, the computed  $p$ -value will always be robust, in the sense that  $p \geq \Pr[Y \geq y]$ .

The structure of the subroutine GENF is given below showing the input and output variables in the algorithm.

SUBROUTINE GENF(R, G, NU, Y, PDFERR, CUMF, LB, UB, NTERMS,  
 EVALS,DENSTY, ERRDEN, IER)

**Table 2. Inputs to GENF**

Text	GENF	Type	Description
$r$	R	integer	df for the numerator of the generalized $F$ distribution with $1 \leq r \leq 25$ where the upper bound is set by MAXP
$\alpha$	G	double precision	weights for the chi-square distributions $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_r > 0$
$\nu$	NU	double precision	df for the denominator of the generalized $F$ distribution with $\nu \geq 1$
$y$	Y	double precision	value for $(T/r)/(V/\nu)$
$PDFERR$	PDFERR	double precision	global truncation error criterion for $pdf$ of $(T/r)/(V/\nu)$ with suggested value = $10^{-4}$

**Table 3. Outputs from GENF**

Text	GENF	Type	Description
$1 - p$	CUMF	double precision	left-hand probability of the $cdf$ of $(T/r)/(V/\nu)$
$LB$	LB	double precision	lower bound for CUMF from Equation 8
$UB$	UB	double precision	upper bound for CUMF from Equation 8
$\tau$	NTERMS	integer	number of terms used in the series expansion
EVALS	EVALS	integer	number of function evaluations required by the adapted integration subroutine
$h_F(w)$	DENSTY	double precision	probability distribution function of $F = (T/r)/(V/\nu)$ at $y$
$\max\{e_\tau(w)\}$	ERRDEN	double precision	maximum local error for $pdf$ of $(T/r)/(V/\nu)$ over values used by integration subroutine
$CSIZE$	CSIZE	double precision	maximum number of partial sum terms $CSIZE = 3000$
$IER$	IER	integer	error code $IER = 1$ if $r < 1$ $IER = 2$ if $r > 10$ $IER = 3$ if weights are not correct $IER = 4$ if $\nu < 1$ $IER = 5$ if $y \leq 0$ $IER = 6$ if $PDFERR \leq 0$ $IER = 7$ if $PDFERR > 0.1$ $IER = 8$ if the value of CSIZE is too small

## References

- [1] Cook, R. (1977). Detection of influential observations in linear regression models, *Technometrics*, 19, 15-18.
- [2] Gurland, J. (1955). Distribution of definite and of indefinite quadratic forms, *Ann. Math. Stat.*, 26, 122-127.
- [3] Hald, A. (1952). *Statistical Theory with Engineering Applications*. John Wiley & Sons, Inc., New York.
- [4] Jensen, D. R. and Ramirez, D. E. (1991). Misspecified  $T^2$  tests. I. Location and scale, *Commun. Statist. -Theory Meth.*, 20, 249-259.
- [5] Jensen, D. R. and Ramirez, D. E. (1998a). Some exact properties of Cook's  $D_I$  statistic. In *Handbook of Statistics*, vol. 16, Balakrishnan, N. and Rao, C. eds., pp. 387-402, Elsevier Science Publishers, Amsterdam.
- [6] Jensen, D. R. and Ramirez, D. E. (1998b). Detecting outliers with Cook's  $D_I$  statistic. *Computing Science and Statistics* 29(1), 581-586.
- [7] Kotz, S., Johnson, N., and Boyd, D. (1967). Series representations of distributions of quadratic forms in normal variables, I: Central case, *Ann. Math. Stat.*, 38, 823-837.
- [8] Longley, H. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user, *J. Amer. Statis. Assoc.*, 62, 819-841.
- [9] Ramirez, D. E. and Jensen, D. R. (1991). Misspecified  $T^2$  tests. II. Series expansions, *Commun. Statist. -Simula.* 20, 97-108.
- [10] Timm, N. (1975). *Multivariate Analysis with Applications in Education and Psychology*. Brooks/Cole Publishing Company, Monterey, California.
- [11] Weisberg, S. (1980). *Applied Linear Regression*. John Wiley & Sons, Inc., New York.