Computation of the Generalized F Distribution

Charles F. Dunkl Donald E. Ramirez
Department of Mathematics
University of Virginia
Charlottesville, VA 22903-3199 USA

Abstract

Exact expressions for the distribution function of a random variable of the form $((\alpha_1\chi^2_{m_1}+\alpha_2\chi^2_{m2})/|m|)/(\chi^2_\nu/\nu)$ are given where the chi-square distributions are independent with degrees of freedom m_1, m_2, and ν respectively. Applications to detecting joint outliers and Hotelling's misspecified T^2 distribution are given.

Key Words: Generalized F distribution, hypergeometric functions, Cook's D_I statistic, outliers, misspecified Hotelling T^2 distribution.

1 Introduction

The *generalized F distribution* is defined as follows. Suppose that the elements of $\mathbf{X} = [\chi^2_{m_1}, \cdots, \chi^2_{m_r}]'$ $(r > 1)$ are independent chi-square random variables with degrees of freedom $(m_1, \cdots, m_r)$, respectively; let $\{\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r > 0\}$ be nonincreasing positive weights; and identify $T = \alpha_1\chi^2_{m_1} + \cdots + \alpha_r\chi^2_{m_r}$. If $\mathcal{L}(V) = \chi^2(\nu)$ independently of $\mathbf{X}$, then the *cdf* of

$$W = \frac{T/|m|}{V/\nu} = \frac{(\alpha_1\chi^2_{m_1} + \cdots + \alpha_r\chi^2_{m_r})/|m|}{V/\nu}, \tag{1}$$

where $|m| = m_1 + \cdots + m_r$, is denoted by $F_r(w; \alpha_1, \cdots, \alpha_r; m_1, \cdots, m_r; \nu)$. If all of the α_i $(1 \leq i \leq r)$ are equal to say α, then the *cdf* of W is denoted by $F_r(w; \alpha; m_1, \cdots, m_r; \nu)$, the scaled central F distribution with degrees of freedom $(|m|, \nu)$. To avoid the trivial case, we will assume that the positive weights are pairwise distinct.

We will give exact expressions for the *pdf* of W for $r = 2$ in terms of the hypergeometric series ${}_2F_1$. This is the analog for generalized functions of the known result for a mixture of two chi-square distributions (Bock and Solomon (1988)). For $r > 2$, we give three numerically tractable expressions for the *pdf* and *cdf* of W. Applications include the detection of joint outliers using Cook's D_I statistics and the calculation of the power of Hotelling's T^2 test with a misspecifed scale.

2 The Distribution of $(T/|m|)/(V/\nu)$

Building on the work of Robbins and Pitman (1949), Gurland (1955), and Kotz, Johnson, and Boyd (1967), Ramirez and Jensen (1991) showed how to compute the *pdf* 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 $\mathcal{L}(V) = \chi^2(\nu - p + 1)$; and they used the notation from Kotz, Johnson and Boyd (1967). We give the results for the general case below where it is convenient for our derivation to use the notation from Robbins and Pitman (1949).

2.1 The Probability Distribution Function for W

Write

$$T = \alpha_r(\frac{\alpha_1}{\alpha_r}\chi^2_{m_1} + \cdots + \frac{\alpha_{r-1}}{\alpha_r}\chi^2_{m_{r-1}} + \chi^2_{m_r}). \tag{2}$$

Following Robbins and Pitman (1949, p. 555) define the constants c_j by the identity

$$A \prod_{i=1}^{r} (1 - u_i z)^{-m_i/2} = \sum_{j=0}^{\infty} c_j z^j, \tag{3}$$

where

$$A = \prod_{i=1}^{r} \left(\frac{\alpha_i}{\alpha_r}\right)^{-m_i/2}. \tag{4}$$

The series in Equation 3 converges absolutely for $|z| < \alpha_1/(\alpha_1 - \alpha_r)$. Set $z = 0$ to see that $c_0 = A$, and set $z = 1$ for the equality $\sum_{j=0}^{\infty} c_j = 1$. Then $P[T \leq y] = \sum c_j G_{|m|+2j}(y/\alpha_r)$, where G_k is the *cdf* for the chi-square distribution with k degrees of freedom. As in Ramirez and Jensen (1991, p. 100), we find that the *pdf* for $W = (T/|m|)/(V/\nu)$ has the representation as stated in the following

Theorem 1 *With the notation above,*

$$\begin{aligned} h_W(w) &= \sum_{j=0}^{\infty} \frac{|m|}{\nu} \frac{c_j}{\alpha_r} \frac{\Gamma\left(\frac{\nu+|m|+2j}{2}\right)\left(\frac{|m|}{\nu}\frac{w}{\alpha_r}\right)^{(|m|+2j-2)/2}}{\Gamma\left(\frac{|m|+2j}{2}\right)\Gamma\left(\frac{\nu}{2}\right)\left(1 + \frac{|m|}{\nu}\frac{w}{\alpha_r}\right)^{(\nu+|m|+2j)/2}} \\ &= \sum_{j=0}^{\infty} \frac{c_j}{\alpha_r} \frac{|m|}{|m|+2j} \; f_F\left(\frac{|m|}{|m|+2j}\frac{w}{\alpha_r}; |m|+2j, \nu\right), \end{aligned} \tag{5}$$

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

A bound for the global truncation error e_τ for the τ^{th} partial sum of the pdf of $W = (T/|m|)/(V/\nu)$ is given by

$$\begin{aligned} & \sum_{j=\tau+1}^{\infty} \frac{c_j}{\alpha_r} \frac{|m|}{|m|+2j}\ f_F\left(\frac{|m|}{|m|+2j}\frac{w}{\alpha_r}; |m|+2j, \nu\right) && (6) \\ \leq\ & \frac{|m|}{\alpha_r(|m|+2(\tau+1))}(1-(c_0+\cdots+c_\tau)) = e_\tau. && (7) \end{aligned}$$

Proof. Use the equality $\sum_{i=0}^{\infty} c_i = 1$, and note that $|f_F(w; v_1, v_2)| \leq 1$ when $v_1 \geq 2$ and $v_2 \geq 1$. ■

The global bound e_τ can be used to determine the number of terms τ to use in the truncated series expansion of the *pdf* for W in Equation 5. In Section 4.2, we improve on the global error bound e_τ by identifying the local error bound as a hypergeometric function ${}_2F_1$.

2.2 Calculation of the Coefficients c_j

Kotz, Johnson, and Boyd (1967) gave the following expression for c_j,

$$\begin{aligned} c_0 &= \prod_{i=1}^{r}\left(\frac{\alpha_r}{\alpha_i}\right)^{m_i/2} = A, \\ d_j &= \sum_{i=1}^{r}\frac{m_i}{2}\left(1-\frac{\alpha_r}{\alpha_i}\right)^j,\ j \geq 1, \qquad (8) \\ c_j &= \frac{1}{j}\sum_{l=0}^{j-1}(d_{j-l}c_l),\ j \geq 1. \end{aligned}$$

We are able to reduce the numerical complexity in the computation of the coefficients c_j by determining a recursive algorithm for c_j. Fix parameters $\mu_1, \ldots, \mu_r$ and variables $u_1, \ldots, u_r$ with $|u_i| < 1$ for all $i\,(1 \leq i \leq r)$. For $k = 0, 1, 2, \ldots$, let

$$P_k = \sum_{|\mathbf{n}|=k}\prod_{i=1}^{r}\frac{(\mu_i)_{n_i}}{n_i!}u_i^{n_i}, \mathbf{n} = (n_1, \ldots, n_r).$$

Note that $\sum_{k=0}^{\infty} P_k = \prod_{i=1}^{r}(1-u_i)^{-\mu_i}$. Denote the set $R = \{1, 2, \ldots, r\}$. For $i \in R$, define

$$\begin{aligned} e_i &= \sum_{S\subset R, |S|=i}\prod_{j\in S}u_j, \\ f_i &= \sum_{S\subset R, |S|=i}\left(\sum_{j\in S}\mu_j\right)\prod_{j\in S}u_j. \end{aligned}$$

Thus e_i is the elementary symmetric function of degree i in $u_1, \dots, u_r$. Then for $k \geq 1$

$$kP_k = \sum_{i=1}^{r} (-1)^{i-1}((k-i)e_i + f_i)P_{k-i}.$$

To prove the identity, let $\lambda_i = \mu_i - 1$ for all i; and for a fixed $\mathbf{n} = (n_1, \dots, n_r)$ with $|\mathbf{n}| = k$, examine the coefficient of $\prod_{i=1}^{r} \frac{(\mu_i)_{n_i}}{n_i!} u_i^{n_i}$ in the sum $kP_k + \sum_{i=1}^{r}(-1)^i((k-i)e_i + f_i)P_{k-i}$. Let $\zeta_i = \frac{n_i}{\mu_i + n_i - 1} = \frac{n_i}{\lambda_i + n_i}$ then this coefficient equals

$$k + \sum_{i=1}^{r} (-1)^i \sum_{S \subset R, |S|=i} (k + \sum_{j \in S} \lambda_j) \prod_{j \in S} \zeta_j.$$

The coefficient of k in this expression is $\prod_{i=1}^{r}(1-\zeta_i)$. For each s, the coefficient of λ_s is

$$\begin{aligned} \zeta_s \sum_{i=1}^{r} (-1)^i \sum \left\{ \prod_{j \in S} \zeta_j : S \subset R \backslash \{s\}, |S| = i-1 \right\} \\ = -\zeta_s \prod_{i \neq s} (1 - \zeta_i). \end{aligned}$$

But $\lambda_s \zeta_s = \frac{\lambda_s n_s}{\lambda_s + n_s} = n_s(1-\zeta_s)$, and so these terms sum to $-\sum_{s=1}^{r} n_s \prod_{i=1}^{r}(1-\zeta_i)$, and $|\mathbf{n}| = k$. This completes the proof by noting that $c_k = AP_k$ with $\mu_i = m_i/2$.

3 Exact Expressions for the *pdf* of W

Use the negative binomial series

$$(1 - sz)^{-b} = \sum_{m=0}^{\infty} s^m \frac{(b)_m}{m!} z^m \tag{9}$$

to express Equation 3 as

$$\sum_{j=0}^{\infty} c_j z^j = A \sum_{j=0}^{\infty} z^j \sum_{i_1 + \cdots + i_{r-1} = j} \prod_{k=1}^{r-1} \frac{u_k^{i_k}}{i_k!} \left(\frac{m_k}{2} \right)_{i_k}, \tag{10}$$

with

$$0 \leq u_i = 1 - \frac{\alpha_r}{\alpha_i} < 1 \, (1 \leq i \leq r).$$

Note that $u_r = 0$. Denote

$$a = \frac{\nu\alpha_r}{|m|}, \tag{11}$$

$$B_0 = a^{\nu/2}\frac{\Gamma\left(\frac{\nu+|m|}{2}\right)}{\Gamma\left(\frac{|m|}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}, \tag{12}$$

$$B_1(w) = \frac{w^{(|m|-2)/2}}{(a+w)^{(\nu+|m|)/2}}, \tag{13}$$

and write the *pdf* for $W = (T/|m|)/(V/\nu)$ with

$$t(w) = \frac{w}{a+w},$$

as

$$h_W(w) = \sum_{j=0}^{\infty}\frac{|m|}{\nu}\frac{c_j}{\alpha_r}\frac{\Gamma\left(\frac{\nu+|m|+2j}{2}\right)\left(\frac{|m|}{\nu}\frac{w}{\alpha_r}\right)^{(|m|+2j-2)/2}}{\Gamma\left(\frac{|m|+2j}{2}\right)\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{|m|}{\nu}\frac{w}{\alpha_r}\right)^{(\nu+|m|+2j)/2}}$$

$$= B_0B_1(w)\sum_{j=0}^{\infty}\frac{c_j\left(\frac{\nu+|m|}{2}\right)_j}{\left(\frac{|m|}{2}\right)_j}t(w)^j \tag{14}$$

$$= AB_0B_1(w)\sum_{j=0}^{\infty}\frac{\left(\frac{\nu+|m|}{2}\right)_j}{\left(\frac{|m|}{2}\right)_j}t(w)^j\sum_{i_1+\cdots+i_{r-1}=j}\prod_{k=1}^{r-1}\frac{u_k^{i_k}}{i_k!}\left(\frac{m_k}{2}\right)_{i_k} \tag{15}$$

$$= AB_0B_1(w)\,F_D^{(r-1)}\left(\frac{\nu+|m|}{2};\frac{m_1}{2},\cdots,\frac{m_{r-1}}{2};\frac{|m|}{2};t(w)u_1,\cdots,t(w)u_{r-1}\right), \tag{16}$$

where F_D is a Lauricella function (Srivastava and Karlsson (1985, p. 41) where we correct the typographical error with Equation 16)). Equation 16 gives a representation of the *pdf* of the distribution W. We will show in Theorem 3 that the *cdf* of W is also a Lauricella $F_D^{(r)}$ function. This representation will yield a numerically computable algorithm for finding p-values. Equation 14 yields a numerically tractable expression for the *pdf* of W. In Section 4.2, we give a tight local truncation error bound $e_\tau(w)$ for determining the number of terms τ to use in the partial sum expression.

3.1 Exact Expressions for the *pdf* of W with $r = 2$

If $r = 2$, Equation 15 is a hypergeometric series, and we have the following result.

Theorem 2 *With the notation above, $a = \nu\alpha_2/|m|$, and $r = 2$, the pdf of W is given by*

$$\begin{aligned} h_W(w) &= AB_0B_1(w)\sum_{j=0}^{\infty}\frac{\left(\frac{\nu+m_1+m_2}{2}\right)_j\left(\frac{m_1}{2}\right)_j}{j!\left(\frac{m_1+m_2}{2}\right)_j}(u_1t(w))^j \quad (17)\\ &= AB_0B_1(w)\,{}_2F_1\left(\begin{matrix}\frac{\nu+m_1+m_2}{2},\frac{m_1}{2}\\ \frac{m_1+m_2}{2}\end{matrix};(1-\frac{\alpha_r}{\alpha_1})\frac{w}{a+w}\right). \quad (18)\end{aligned}$$

To find the *cdf* of W when $r = 2$, integrate $h_W(w)$ in Equation 18.

We note that if we had used the notation of Kotz, Johnson, and Boyd (1967) and scaled y by y/δ with $0 < \delta < \alpha_r$, then $u_2 > 0$. In this situation, we would use the Bailey transformation (Srivastava and Karlsson (1985, p. 304)) to convert the two variable hypergeometric series in Equation 17 to the ${}_2F_1$ function in Equation 18.

3.2 Exact Expressions for the *cdf* of W with $r \geq 2$

The Lauricella function $F_D^{(r-1)}$ in Equation 16 has an integral representation (Exton, 1976, p. 49) where the domain of integration is over the simplex E_r with $x_1 + \cdots + x_r = 1\,(x_i \geq 0, 1 \leq i \leq r)$ as

$$\begin{aligned} &F_D^{(r-1)}\left(\frac{\nu+|m|}{2};\frac{m_1}{2},\cdots,\frac{m_{r-1}}{2};\frac{|m|}{2};t(w)u_1,\cdots,t(w)u_{r-1}\right)\\ &= \Gamma\left[\frac{\frac{|m|}{2}}{\frac{m_1}{2},\cdots,\frac{m_r}{2}}\right]\int_{E_r}(1-\sum_{i=1}^{r-1}t(w)u_ix_i)^{-\frac{\nu+|m|}{2}}\prod_{i=1}^{r}x_i^{\frac{m_i}{2}-1}d\mathbf{x}. \quad (19)\end{aligned}$$

In Dunkl and Ramirez (1994a, 1994b), we computed the surface measure of ellipsoids using hyperelliptic integrals. We showed that the $(n-1)$-dimensional hyperelliptic integral could be transformed into a univariate integral using the Euler integral representation (Exton, 1976, p. 49) for F_D. This transformation does not apply to Equation 19 since $\frac{\nu+|m|}{2} > \frac{|m|}{2}$. Here we will use a different approach.

We show how to represent the *cdf* of the generalized F distribution W as a Lauricella $F_D^{(r)}$ function. This representation will provide a numerically tractable procedure for computing the *cdf* of W, denoted by $H_W(w)$, which does not require integrating the *pdf* of W.

Theorem 3 *With the notation above and $r \geq 2$, the cdf of W is given by*

$$\begin{aligned} &H_W(y) = AB_0\frac{y^{|m|/2}}{(|m|/2)(a+y)^{(\nu+|m|)/2}}\\ &F_D^{(r)}(\frac{\nu+|m|}{2};\frac{m_1}{2},\cdots,\frac{m_{r-1}}{2},1;\frac{|m|}{2}+1;t(y)u_1,\cdots,t(y)u_{r-1},t(y)), \quad (20)\end{aligned}$$

with $a = \nu\alpha_r/|m|$ and $t(y) = y/(a+y)$ as before.

Proof. From Equations 16 and 19, write the *cdf* of W as

$$H_W(y) = \int_0^y h_W(w)dw$$

$$= AB_0 \frac{\Gamma\left(\frac{|m|}{2}\right)}{\prod_{i=1}^r \Gamma\left(\frac{m_i}{2}\right)} \int_0^y \frac{w^{|m|/2-1}}{(a+w)^{(\nu+|m|)/2}}$$

$$\int_{E_r} \left(1 - \sum_{i=1}^r \frac{w}{a+w} u_i x_i\right)^{-(\nu+|m|)/2} \prod_{i=1}^r x_i^{m_i/2-1} d\mathbf{x}dw$$

$$= AB_0 \frac{\Gamma\left(\frac{|m|}{2}\right)}{\prod_{i=1}^r \Gamma\left(\frac{m_i}{2}\right)} \int_0^y \int_{E_r} w^{r-1}(a + w - \sum_{i=1}^r w u_i x_i)^{-(\nu+|m|)/2}$$

$$\prod_{i=1}^r (w x_i)^{m_i/2-1} d\mathbf{x}dw.$$

Change variables with $s_i = wx_i/y$ $(1 \le i \le r)$ and $s_{r+1} = 1 - w/y$. Note that $\sum_{i=1}^r s_i = w/y$ with the absolute value of the inverse Jacobian $J^{-1} = \left|\frac{\partial(x_1,\cdots,x_{r-1},w)}{\partial(s_1,\cdots,s_{r-1},s_r)}\right| = w^{r-1}/y^r$. Thus

$$H_W(y) = AB_0 \frac{\Gamma\left(\frac{|m|}{2}\right)}{\prod_{i=1}^r \Gamma\left(\frac{m_i}{2}\right)}$$

$$\int_{E_{r+1}} y^r \left(a + y(1 - s_{r+1}) - \sum_{i=1}^r y s_i u_i\right)^{-(\nu+|m|)/2} \prod_{i=1}^r (y s_i)^{m_i/2-1} d\mathbf{s}$$

$$= AB_0 \frac{\Gamma\left(\frac{|m|}{2}\right)}{\prod_{i=1}^r \Gamma\left(\frac{m_i}{2}\right)} (a+y)^{-(\nu+|m|)/2} y^{|m|/2} dw$$

$$\int_{E_{r+1}} \left(1 - \frac{y s_{r+1}}{a+y} - \sum_{i=1}^r \frac{y}{a+y} s_i u_i\right)^{-(\nu+|m|)/2} \prod_{i=1}^r s_i^{m_i/2-1} d\mathbf{s}$$

$$= AB_0 \frac{1}{|m|/2} \frac{y^{|m|/2}}{(a+y)^{(\nu+|m|)/2}} \tag{21}$$

$$F_D^{(r)}\left(\frac{\nu+|m|}{2}; \frac{m_1}{2}, \cdots, \frac{m_{r-1}}{2}, 1; \frac{|m|}{2} + 1; t(y)u_1, \cdots, t(y)u_{r-1}, t(y)\right),$$

with $a = \nu\alpha_r/|m|$ and $t(y) = y/(a+y)$. ∎

To convert Equation 21 into a numerically tractable series, write

$$
\begin{aligned}
H_W(y) &= AB_0 \frac{1}{|m|/2} \frac{y^{|m|/2}}{(a+y)^{(\nu+|m|)/2}} \\
& F_D^{(r)}\left(\frac{\nu+|m|}{2}; \frac{m_1}{2}, \cdots, \frac{m_{r-1}}{2}, 1; \frac{|m|}{2}+1; t(y)u_1, \cdots, t(y)u_{r-1}, t(y)\right) \\
&= B_0 \frac{y^{|m|/2}}{(a+y)^{(\nu+|m|)/2}} \\
& \sum_{j=0}^{\infty} \frac{(\frac{\nu+|m|)}{2})_j}{(\frac{|m|}{2})_{j+1}} \left(\frac{y}{a+y}\right)^j \left[A \sum_{i_1+\cdots+i_{r-1}\leq j} \prod_{k=1}^{r-1} \frac{u_k^{i_k}}{i_k!} \left(\frac{m_k}{2}\right)_{i_k}\right] \\
&= B_0 \frac{y^{|m|/2}}{(a+y)^{(\nu+|m|)/2}} \sum_{j=0}^{\infty} \frac{(\frac{\nu+|m|)}{2})_j}{(\frac{|m|}{2})_{j+1}} \left(\frac{y}{a+y}\right)^j \sum_{i=0}^{j} c_i. \qquad (22)
\end{aligned}
$$

4 Local Truncation Error Bounds

Denote by $\widehat{h}_W(w)$ and $\widehat{H}_W(y)$ the partial sum estimates for $h_W(w)$ and $H_W(y)$, respectively, from Equations 14 and 23. In this Section, we derive local truncation error bounds to determine the number of terms required by the partial sums.

4.1 Local Truncation Error Bound $e_\tau^*(y)$ for the *cdf* of W

For Equation 22 to be numerically tractable, we derive the local truncation error. Write $t(y) = (y/(a+y)) < 1$,

$$
\begin{aligned}
H_W(y) &= B_0 y B_1(y) \frac{y}{(|m|/2)} \sum_{j=0}^{\infty} \frac{(\frac{\nu+|m|)}{2})_j}{(\frac{|m|}{2}+1)_j} t(y)^j \sum_{i=0}^{j} c_i \\
&= B_0 B_1(y) \frac{y}{(|m|/2)} \sum_{j=0}^{\tau} \frac{(\frac{\nu+|m|)}{2})_j}{(\frac{|m|}{2}+1)_j} t(y)^j \sum_{i=0}^{j} c_i + \qquad (23) \\
& B_0 B_1(y) \frac{y}{(|m|/2)} \sum_{j=\tau+1}^{\infty} \frac{(\frac{\nu+|m|)}{2})_j}{(\frac{|m|}{2}+1)_j} t(y)^j \sum_{i=0}^{j} c_i
\end{aligned}
$$

$$= \widehat{H}_W(y) + B_0 B_1(y) \frac{y}{(|m|/2)} \frac{(\frac{\nu+|m|)}{2})_{\tau+1}}{(\frac{|m|}{2}+1)_{\tau+1}} t(y)^{\tau+1} \times$$

$$\sum_{j=0}^{\infty} \frac{(\frac{\nu+|m|)}{2}+\tau+1)_j}{(\frac{|m|}{2}+\tau+2)_j} t(y)^j (1 - 1 + \sum_{i=0}^{\tau+1+j} c_i) \tag{24}$$

$$= \widehat{H}_W(y) + B_0 B_1(y) \frac{y}{(|m|/2)} \frac{(\frac{\nu+|m|)}{2})_{\tau+1}}{(\frac{|m|}{2}+1)_{\tau+1}} t(y)^{\tau+1} {}_2F_1 \left(\begin{matrix} \frac{\nu+|m|)}{2}+\tau+1, 1 \\ \frac{|m|}{2}+\tau+2 \end{matrix} ; t(y) \right)$$

$$-B_0 B_1(y) \frac{y}{(|m|/2)} \frac{(\frac{\nu+|m|)}{2})_{\tau+1}}{(\frac{|m|}{2}+1)_{\tau+1}} t(y)^{\tau+1} \sum_{j=0}^{\infty} \frac{(\frac{\nu+|m|)}{2})_j}{(\frac{|m|}{2}+1)_j} t(y)^j (1 - \sum_{i=0}^{\tau+1+j} c_i).$$

The partial sum estimate $\widehat{H}_W(y)$ can be enhanced by identifying most of the truncation error as a scaled ${}_2F_1$ hypergeometric functions. The remaining truncation error is bounded by a scaled ${}_2F_1$ function and is stated in the following.

Theorem 4 *With the notation above, the estimated $P[W \leq y]$ is given by*

$$\widehat{H}_W(y) + B_0 B_1(y) \frac{y}{(|m|/2)} \frac{(\frac{\nu+|m|)}{2})_{\tau+1}}{(\frac{|m|}{2}+1)_{\tau+1}} t(y)^{\tau+1} {}_2F_1 \left(\begin{matrix} \frac{\nu+|m|)}{2}+\tau+1, 1 \\ \frac{|m|}{2}+\tau+2 \end{matrix} ; t(y) \right) \tag{25}$$

with local truncation error bound given by

$$e_\tau^*(y) = (1 - \sum_{i=0}^{\tau+1} c_i) B_0 B_1(y) \frac{y}{(|m|/2)} \frac{(\frac{\nu+|m|)}{2})_{\tau+1}}{(\frac{|m|}{2}+1)_{\tau+1}} t(y)^{\tau+1} \times \tag{26}$$

$${}_2F_1 \left(\begin{matrix} \frac{\nu+|m|)}{2}+\tau+1, 1 \\ \frac{|m|}{2}+\tau+2 \end{matrix} ; t(y) \right).$$

To find τ, we increase the size of τ unless the remaining error $e_\tau^*(y)$ from Equation 26 is less than a prescribed small value. The suggested value is 10^{-4}.

4.2 Local Truncation Error Bound $e_\tau(w)$ for the *pdf* of W

Recall that Equation 14 yields a numerically tractable expression for the *pdf* of W. A tight local truncation error bound for determining the number of terms τ to use in the partial sum expression follows as above and is stated in the following.

Theorem 5 *With the notation above,*

$$\widehat{h}_W(w) = B_0 B_1(w) \sum_{j=0}^{\tau} \frac{c_j \left(\frac{\nu+|m|}{2}\right)_j}{\left(\frac{|m|}{2}\right)_j} t(w)^j \tag{27}$$

with local truncation error bound given by

$$e_\tau(w) = c_{\tau+1} B_0 B_1(w) \frac{(\frac{\nu+|m|)}{2})_{\tau+1}}{(\frac{|m|}{2})_{\tau+1}} t(w)^{\tau+1} \, {}_2F_1 \left(\begin{matrix} \frac{\nu+|m|)}{2} + \tau + 1, 1 \\ \frac{|m|}{2} + \tau + 2 \end{matrix} ; t(w) \right), \tag{28}$$

a scaled ${}_2F_1$ hypergeometric function.

To determine the number of terms for the partial sum estimate $\widehat{h}_W(w)$, increase the size of τ unless the local truncation error from Equation 28 is less than a prescribed small value. The suggested value is $y10^{-4}$ where the p-value is calculated from y.

5 Applications

We will give two applications where the distribution of the test statistic is the generalized F distribution.

5.1 Detection of Outliers

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

$$\mathbf{Y}_0 = \mathbf{X}_0 \boldsymbol{\beta} + \boldsymbol{\varepsilon}_0, \tag{29}$$

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, $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of unknown parameters, and $\boldsymbol{\varepsilon}_0$ is a $(N \times 1)$ vector of randomly distributed Gaussian errors with $E(\boldsymbol{\varepsilon}_0) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}_0) = \sigma^2 \mathbf{I}_N$. The least squares estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}_0^{'} \mathbf{X}_0)^{-1} \mathbf{X}_0^{'} \mathbf{Y}_0$. 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 $\boldsymbol{\beta}$? Cook's D_I statistics are based on a Mahalanobis distance between $\hat{\boldsymbol{\beta}}$ (using all the cases) and $\hat{\boldsymbol{\beta}}_I$ (using all cases except those in the subset I), as given by

$$D_I(\hat{\boldsymbol{\beta}}, \mathbf{M}, c\hat{\sigma}^2) = (\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})^{'} \mathbf{M} (\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}}) / (c\hat{\sigma}^2), \tag{30}$$

with a $(k \times k)$ nonnegative definite matrix M, $\hat{\sigma}^2$ is an unbiased estimate of the variance, and a user defined constant c. We use $c = r$ and the estimator s_I^2, the sample variance estimator with the cases in I omitted We will discuss the case with $\mathbf{M} = \mathbf{X}^{'} \mathbf{X}$, where $\mathbf{X}$ denotes the remaining rows of $\mathbf{X}_0$. We have chosen s_I^2 as the estimator for σ^2 since this estimator and the numerator of Equation 30 are independent.

Using the results in this paper, we are able to numerically compute the *cdf* of Cook's D_I statistics in the case of joint outliers, 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.

5.1.1 Notation

To fix the notation, let I be a subset of $\{1, \ldots, N\}$, say $I = \{i_1, \ldots, i_r\}$. 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 $\boldsymbol{\varepsilon}_0^{'} = [\boldsymbol{\varepsilon}_1^{'}, \boldsymbol{\varepsilon}_2^{'}]$. Thus Equation 29 has been transformed into

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Z} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}. \tag{31}$$

The ordered eigenvalues of $\mathbf{Z}(\mathbf{X}_0^{'}\mathbf{X}_0)^{-1}\mathbf{Z}^{'}$ are denoted $\{\lambda_1 \geq \cdots \geq \lambda_r > 0\}$ usually called the canonical leverages. 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

$$F_r(w\,; \alpha_1; \nu) \leq F_r(w\,; \alpha_1, \ldots, \alpha_r; 1, \ldots, 1; \nu) \leq F_r(w\,; \alpha^*; \nu)\ , \tag{32}$$

with the maximum weight α_1, the geometric mean α^* of the weights $\{\alpha_1, \ldots, \alpha_r\}$, and $F_r(w\,; \alpha; \nu)$ the scaled central F distribution.

The basic characterization theorem for D_I is given in Jensen and Ramirez (1998a) and is:

Theorem 6 *Suppose that* $\mathcal{L}(\mathbf{Y}) = N_N(\mathbf{X}_0\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)$*, then the distribution of* $D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}^{'}\mathbf{X}, rs_I^2)$ *is given by* $F_r(w; \lambda_1, \cdots, \lambda_r; 1, \cdots, 1; N - r - k)$.

With $r = 1$, $\mathcal{L}(D_i(\hat{\boldsymbol{\beta}}, \mathbf{X}^{'}\mathbf{X}, s_i^2)/\lambda_1) = F(1, N - 1 - k)$. Outliers also can be tested using the studentized deleted residuals with $\mathcal{L}((y_i - \hat{y}_{(i)})/(s_i(1 + \mathbf{x}_i(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{x}_i^{'})^{1/2})) = t(N - 1 - k)$ 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 λ_1. In Jensen and Ramirez (1998b) it is shown that the p-values from these two tests are also equal to the p-values from Theorem 6. Thus, in case of single deletion with $r = 1$, all of these three standard tests for outliers will have a common p-value.

5.1.2 Examples

For the Hald (1952, p. 647) data set ($N = 13$ and $k = 5$) using the test statistic $D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}^{'}\mathbf{X}, 2s_I^2)$ and the global bounds in Equation 32, 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 $0.01305 < p_I < 0.04610$. Using $m_1 = m_2 = 1$, the canonical leverages $\boldsymbol{\lambda} = (0.408676, 0.124019)$ for the weights $\boldsymbol{\alpha}$, the degrees of freedom $\nu = N - r - k = 6$, and the observed Cook's D_I statistic $y = 2.19331$, we can now easily compute from Equation 18 that the p-value is p_I= 0.02181.

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 the test statistic $D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}'\mathbf{X}, 2s_I^2)$, with $r = 2$, the canonical leverages $\boldsymbol{\lambda} = (0.690029, 0.614130)$ for the weights, $\nu = N - r - k = 16 - 2 - 7 = 7$, and the observed Cook's D_I statistic $y = 1.812433$, we compute that the p-value is $p_I = 0.12927$.

Using the test statistic $D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}'\mathbf{X}, 2s_I^2)$ and the global bounds Equation 32, 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 $\boldsymbol{\lambda} = (0.615959, 0.371827)$, $y = 2.57861$, and $0.03822 \leq p_{I_1} = 0.04186 \leq 0.06356$, (2) $I_2 = \{4, 15\}$ with $\boldsymbol{\lambda} = (0.505387.0.393672)$, $y = 1.76885$, and $0.04961 \leq p_{I_2} = 0.04982 \leq 0.05555$, and (3) $I_3 = \{10, 16\}$ with $\boldsymbol{\lambda} = (0.736874, 0.695572)$, $y = 2.57906$, and $0.03761 \leq p_{I_3} = 0.04571 \leq 0.07979$ where the p-values p_I are computed from Equation 18.

Our recommendation to the practitioner, who wishes to find joint outliers, is to initially screen for potential joint outliers using Equation 32 with $D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}'\mathbf{X}, rs_I^2)$. If $r = 1$ then the distribution of D_i is a scaled central F distribution. If $r = 2$ then the distribution of D_I is a scaled ${}_2F_1$ series. If $r > 2$ then use Equation26 to find the numbers of terms required to have the local truncation error small. The suggested value for the bound is 10^{-4}. The p-values for the *cdf* for the distribution of $D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}'\mathbf{X}, rs_I^2)$ are calculated using the enhanced truncated series in Equation 25.

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

Theorem 7 *The distribution of the test statistic* $((\nu^* - p + 1)/p)(T^2/\nu^*)$ *is the generalized* F *distribution* $F_r(w; \pi_1, \cdots, \pi_p; 1, \cdots, 1; \nu^* - p + 1)$.

5.2.1 Hotelling's misspecifed scale distribution

The conventional model for T^2 is based on a random sample $\{X_1, \ldots, X_N\}$ from $N_p(\boldsymbol{\mu}, \Sigma)$ using the unbiased sample means and dispersion matrix $(\bar{\mathbf{X}}, \mathbf{S})$. We have $\mathcal{L}(\bar{\mathbf{X}}) = \mathbf{N}_p(\boldsymbol{\mu}, \frac{1}{N}\Sigma)$ 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 $T^2 = (N-1)(\bar{\mathbf{X}} - \boldsymbol{\mu})'(\frac{N-1}{N}\mathbf{S})^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) = N_p(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ and $\mathcal{L}(((N-p)/p)(T^2/(N-1))) = F(p, N-p)$, the central F distribution when $N > p$. If the process dispersion parameters have shifted, then 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 $F_r(w; \pi_1, \cdots, \pi_p; 1, \cdots, 1; N-p)$. Here $r = p$, $\nu = \nu^* - p + 1 = N - p$, and $\{\pi_1 \geq \pi_2 \geq \cdots \geq \pi_p > 0\}$ the ordered roots of $\Omega^{-\frac{1}{2}}\Sigma\Omega^{-\frac{1}{2}}$.

5.2.2 Examples

An important application of generalized F distributions is for computing the power of a misspecified Hotelling's T^2 test for a multivariate quality control chart. Power analysis for a misspecified mean $\boldsymbol{\mu}$ is standard. Using generalized F distributions, the power analysis for a misspecified covariance Ω can be performed. If a process changes, not only will the mean change but generally the covariance structure will also change. The robustness of T^2 under misspecification of scale can be verified by computing the cumulative density of T^2 for varying choices of $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_p > 0$ at the critical value of T^2. For example, if Ω_ρ is a 3×3 equicorrelated matrix ($r = p = 3$) with $\rho = 0.5$, and if Σ 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 $P[Y = ((N-p)/p)(T^2/(N-1)) \geq 3.8625] = 0.12310$. In this example, $\pi_1 = \pi_2$, so we could compute the p-values exactly from Theorem 1, with $F_3(w; \pi_1, \pi_2, \pi_3; 1, 1, 1; N-p) = F_2(w; \pi_1, \pi_3; 2, 1; N-p)$. Instead, we use this problem to demonstrate the number of terms required by the three numerical methods discussed in this paper.

In Table 1, we present similar computations for varying ρ. For each ρ 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 give the value of $P[Y = ((N-p)/p)(T^2/(N-1)) \geq 3.8625$. Also shown are the number of terms required using the three numerical presented in this paper. The first is τ_1 from Equation 7 required to satisfy $ye_{\tau_1} \leq 10^{-4}$, the second is τ_2 from Equation 28 required to satisfy $ye_{\tau_2}(y) \leq 10^{-4}$, and the third is τ_3 from Equation 26 required to satisfy $e^*_{\tau_2}(y) \leq 10^{-4}$. 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

ρ	τ_1	τ_2	τ_3	$P[Y \geq 3.8625]$
0.0	1	1	1	0.0500
0.1	6	7	6	0.0526
0.2	10	11	8	0.0600
0.3	15	15	12	0.0727
0.4	20	20	16	0.0926
0.5	28	26	21	0.1231
0.6	40	32	27	0.1704
0.7	58	40	34	0.2458
0.8	92	49	43	0.3712
0.9	185	58	55	0.59055

As anticipated, the numbers of terms τ required is fewer when the enhanced partial sum from Equation 25 is used. More importantly, the method from Section 4.1 does not require that the *pdf* to be numerically integrated.

6 Conclusion

We have derived the exact distribution of the generalized F distribution $F_2(w; \alpha_1, \alpha_2; m_1, m_2; 2)$ in terms of the hypergeometric series ${}_2F_1$. This extends the corresponding result of Bock and Solomon for a mixture of two chi-square distributions to the generalized F distribution with $r = 2$. Explicit representations for the case $r \geq 2$ are given in terms of a Lauricella F_D functions. Numerically computable series expansion have been derived. Applications to the detection of joint outliers and to the misspecified Hotelling T^2 statistic have been given.

References

[1] Bock, M. and Solomon, H. (1988). Distributions of quadratic forms. *The Australian Journal of Statistics* 3**0A**, 139-149.

[2] Cook, R. (1977). Detection of influential observations in linear regression models. *Technometrics* **19**, 15-18.

[3] Dunkl, C. F. and Ramirez, D. E. (1994a). Computing hyperelliptic integrals for surface measure of ellipsoids. *Transactions on Mathematical Software* **20(4)**, 413-426.

[4] Dunkl, C. F. and Ramirez, D. E. (1994b). Algorithm 736: Hyperelliptic integrals and the surface measure of ellipsoids. *Transactions on Mathematical Software* **20(4)**, 427-435.

[5] Exton, H. (1976). *Multiple Hypergeometric Functions and Applications.* John Wiley & Sons, Inc., New York.

[6] Gurland, J. (1955). Distribution of definite and of indefinite quadratic forms. *The Annals of Mathematical Statistics* **26**, 122-127.

[7] Hald, A. (1952). *Statistical Theory with Engineering Applications*. John Wiley & Sons, Inc., New York.

[8] Jensen, D. R. and Ramirez, D. E. (1991). Misspecified T^2 tests. I. Location and scale. *Communications in Statistics A, Theory and Methods* **20**, 249-259.

[9] 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.

[10] Jensen, D. R. and Ramirez, D. E. (1998b). Detecting outliers with Cook's D_I statistic. *Computing Science and Statistics* **29(1)**, 581-586.

[11] Kotz, S., Johnson, N., and Boyd, D. (1967). Series representations of distributions of quadratic forms in normal variables, I: Central case. *The Annals of Mathematical Statistics* **38**, 823-837.

[12] Longley, H. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association* **62**, 819-841.

[13] Ramirez, D. E. and Jensen, D. R. (1991). Misspecified T^2 tests. II. Series expansions. *Communications in Statistics B, Simulation and Computation* **20**, 97-108.

[14] Robbins, H. and Pitman, E. (1949). Application of the method of mixtures to quadratic forms in normal variates. *The Annals of Mathematics Statistics* **20** 552-560.

[15] Srivastava, H. and Karlsson, P. (1985). *Multiple Gaussian Hypergeometric Series*. Ellis Horwood Limited, Chichester, England.