A SAS\textsuperscript{R} MATRIX SOFTWARE MACRO FOR TESTING THE RANDOM EFFECTS IN RANDOM MODELS WITH UNEQUAL CELL FREQUENCIES IN THE LAST STAGE

José Gallo, Shands Hospital at the University of Florida  
Richard MacElderry, Shands Hospital at the University of Florida  
André Khuri, University of Florida

Abstract

The testing of both the fixed effects and random effects in unbalanced mixed models (which include the random-effects models) has relied on approximate procedures such as Satterthwaite's approximation. These approximate methods are often unreliable as discussed in Tietjen (1974) and Cummings and Gayler (1974).

Khuri (1989) derived exact tests concerning the variance components for general random models that are unbalanced with respect to the last stage of their associated designs. Khuri's article represents a generalization of the results introduced earlier by Khuri and Littell (1987) with regard to the unbalanced random two-way classification model. An outline of Khuri's (1989) method is presented in Section 2. A SAS software macro based on this method has been developed and is briefly described in Section 3.

1. Introduction

A number of exact tests concerning variance components have been proposed in earlier literature. Approaches were suggested in a series of papers by Wald (1940, 1941, and 1947). He suggested ways of creating quadratic forms that are monotonic functions of the parameters of interest. Later, Spjøtvoll (1968) and Thomsen (1975) derived exact tests for the unbalanced random two-way classification model. In particular, Thomsen's paper was a background for the approach used by Khuri and Littell (1987) with the two-factor random model and later generalized to more complex models (Khuri, 1989).

Thomsen (1975) noted that the variance-covariance matrix of the cell means for the two-way cross-classification model could be written in matrix form as

\[ \text{Var}(\gamma) = A_1 \sigma_\alpha^2 + A_2 \sigma_\beta^2 + I \sigma_{\alpha \beta}^2 + K \sigma^2, \]

where \( A_1 \) and \( A_2 \) are associated with the design matrices of the main effects: \( \alpha \) and \( \beta \), whose variance components are \( \sigma_\alpha^2 \) and \( \sigma_\beta^2 \), respectively. The matrix \( K \) is a diagonal matrix with diagonal elements equal to the reciprocal of the cell frequencies.

Since \( A_1 \) and \( A_2 \) commute there exists an orthogonal matrix \( P \) that diagonalizes them simultaneously. Thomsen (1975) considered the transformation \( \gamma = P \zeta \).

This leads to the following variance-covariance matrix:

\[ \text{Var}(\zeta) = PA_1 \sigma_\alpha^2 P' + PA_2 \sigma_\beta^2 P' + I \sigma_{\alpha \beta}^2 P' + PKP' \sigma^2 \]

\[ = D_1 \sigma_\alpha^2 + D_2 \sigma_\beta^2 + I \sigma_{\alpha \beta}^2 + PKP' \sigma^2, \]

where \( D_1 \) and \( D_2 \) are diagonal matrices whose diagonal elements are the eigenvalues of \( A_1 \) and \( A_2 \), respectively. Since the number of nonzero eigenvalues of a symmetric matrix is equal to its rank, and (i) rank \( (A_1) = s \); (ii) rank \( (A_2) = r \); and (iii) rank \( (A_1 + A_2) = r + s - 1 \); it is possible to partition the vector \( \gamma \) into subvectors whose variance-covariance matrices do not depend on both \( \sigma_\alpha^2 \) and \( \sigma_\beta^2 \). However, they depend on the variance components associated with the interaction and error terms, \( \sigma_{\alpha \beta}^2 \) and \( \sigma^2 \), respectively.

Thomsen (1975) was then able to obtain exact tests for the variance components associated with the main effects, \( \alpha \) and \( \beta \), provided that \( \sigma_{\alpha \beta}^2 = 0 \) (The exact tests derived by Spjøtvoll (1968) have the same requirement).

Khuri and Littell (1987) considered an additional step in the process of diagonalizing the variance-covariance matrix of \( \gamma \) based on the eigenvalues associated with the matrix \( PKP' \). Their exact tests for the variance components associated with the main effects do not require that \( \sigma_{\alpha \beta}^2 = 0 \). The only assumption for these tests is that \( n_x > 2s + 1 \), where \( n_x \) is the sample size and \( s \) is the product of the levels associated with the two factors in the model.

In Section 2, we provide an outline of a more general approach by Khuri (1989) that applies to a general random model having imbalance only in the last stage. In Section 3, we consider a SAS software macro based on this approach.


Consider the following model:

\[ \gamma = \sum_{i=0}^{n} X_i \beta_i + \epsilon, \]

where \( \beta_i \) are independent normally distributed random vectors with mean \( \mathbf{0} \) and variance \( \sigma_i^2 \), and furthermore, independently distributed from \( \epsilon \). The vector \( \gamma \) is \( \text{MVN}(\mathbf{0}, \sigma^2 I) \). The matrix \( X_i \) is a matrix of zeros and ones of order \( n \times 1 \), \( (X_0 \beta_0) \) is an \( n \times 1 \) column vector of ones and \( \epsilon \) represents the number of elements in \( \beta_i \).
If we write the model based on the cell means rather than the individual observations, we obtain the vector $\mathbf{y}$ which we can express in matrix notation as

$$\mathbf{y} = \sum_{i=0}^{c} \mathbf{H}_i \mathbf{b}_i + \mathbf{r},$$

(2.2)

where $\mathbf{H}_i$ is a matrix involving the direct product of identity matrices and vectors of ones.

Note that the variance-covariance matrix of $\mathbf{y}$ can be expressed as

$$\text{Var}(\mathbf{y}) = \sum_{i=1}^{c} A_i \sigma_i^2 + \mathbf{K},$$

(2.3)

where $A_i = \mathbf{H}_i \mathbf{H}_i'$ and $\mathbf{K}$ represents a diagonal matrix whose diagonal element is the reciprocal of the corresponding cell frequency.

Note that, in general, it is not possible to express $\text{Var}(\mathbf{y})$, such that the variance components are displayed in a fashion that will produce exact tests. For this reason the variance-covariance matrix must be diagonalized.

Consider the following result from matrix theory: If the $A_i$'s commute, that is, $A_i A_j = A_j A_i$, then there exists an orthogonal matrix $\mathbf{Q}$ that simultaneously diagonalizes the $A_i$'s. The matrix $\mathbf{Q}$ can be obtained as follows:

1. Let $\mathbf{P}_i$ be the matrix associated with the sum of squares for the $i$th effect ($i=0, 1, \ldots, \nu$) for model (2.2) having rank $m_i$. This model is treated as a balanced model with one observation per cell.
2. Let $\mathbf{Q}_i$ be the Gram-Schmidt orthonormal factorization of the rows (or columns) of $\mathbf{P}_i$.
3. Let $\mathbf{Q}$ be the matrix defined by

$$\mathbf{Q} = \left[ \mathbf{Q}_0' : \mathbf{Q}_1' : \ldots : \mathbf{Q}_\nu' \right].$$

(2.4)

From Khuri (1969, Lemma 3.4 (ii), (iii)),

$$\mathbf{Q}_i \mathbf{Q}_j' = \begin{cases} 1_{m_i} & i = j = 0, 1, \ldots, m_i \\ 0 & i \neq j \end{cases}$$

(2.5)

and

$$A_j \mathbf{Q}_i' = \begin{cases} 0 & \text{if the subscripts associated with the } i\text{-th effect are not contained with the set of subscripts for the } j\text{-th effect.} \\ b_j \mathbf{Q}_i' & \text{if the subscripts associated with the } i\text{-th effect are also subscripts for the } j\text{-th effect.} \end{cases}$$

(2.6)

where $b_j = c/c_j$, and $c$ equals the number of cells and $c_j$ is defined in (2.1). It follows that $\text{Var}(\mathbf{Q}_i \mathbf{y})$ can be written as

$$\text{Var}(\mathbf{Q}_i \mathbf{y}) = r_j \mathbf{m}_j + \sigma^2 (\mathbf{Q}_i \mathbf{K} \mathbf{Q}_i'),$$

(2.7)

where $r_j$ is a linear combination of the variance components of those effects whose set of subscripts contain the subscripts associated with the $i$th effect.

If we let $\mathbf{y} = \mathbf{Q} \mathbf{x}$, where $\mathbf{Q} = \left[ \mathbf{Q}_0' : \mathbf{Q}_1' : \ldots : \mathbf{Q}_\nu' \right]$, we have from (2.7) that

$$\text{Var}(\mathbf{y}) = \text{Diag}(r_1 \mathbf{1}_{m_1}, r_2 \mathbf{1}_{m_2}, \ldots, r_\nu \mathbf{1}_{m_\nu}) + \sigma^2 \mathbf{Q} \mathbf{K} \mathbf{Q}'.$$

(2.8)

Since $\mathbf{Q}_0$ is not included, we also have that $\mathbf{E}(\mathbf{y}) = \mathbf{0}$.

Note that our variance-covariance matrix is diagonal except for the contribution made by the error term $\mathbf{Q} \mathbf{K} \mathbf{Q}$$'$.

The following steps lead to a completely diagonal variance-covariance matrix.

We can express the error sum of squares as

$$\text{SSE} = \mathbf{y} \mathbf{R} \mathbf{y}',$$

(2.9)

where $\mathbf{R} = \mathbf{I} - \mathbf{W} \mathbf{W}'$ and $\mathbf{W} = \text{Diag}(1_{m_1})$.

$m_1$ represents the frequency of cell $t$ and each diagonal position of $\mathbf{W}$ corresponds to a cell.

The matrix $\mathbf{R}$ is idempotent of rank $n-c$, where $c$ is the product of the ranges of the subscripts associated with the classification variables in the model. (Hence, if there are only two factors in the model indexed by $i$ and $j$ with $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, b$ then $c = ab$). Therefore, $\mathbf{R}$ has eigenvalues that are either 0 or 1. As a result, there exists an orthogonal matrix $\mathbf{H}$ such that

$$\mathbf{R} = \mathbf{H} \mathbf{D} \mathbf{H}'$$

(2.10)

where $\mathbf{D}$ is a diagonal matrix having $n-c$ diagonal elements equal to unity and $c$ diagonal elements equal to zero.
We can partition this matrix into

\[ R = \begin{bmatrix} \mathbf{I}_1 & \mathbf{I}_2 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I}_1 \mathbf{H}_1 \mathbf{I}_1 & \mathbf{I}_1 \mathbf{H}_2 \mathbf{I}_2 & \mathbf{0} \end{bmatrix} \]  \tag{2.11}

where \( \mathbf{I}_1 = \mathbf{I} - \mathbf{E} \), \( \mathbf{I}_2 = \mathbf{N} - \mathbf{C} \), and \( \mathbf{0} \) is a zero matrix of order \( c \times c \). Note that we must have \( n-2c+1>0 \).

The matrix \( \mathbf{H} \) is similarly partitioned into

\[ \mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 & \mathbf{H}_3 \end{bmatrix} \]  \tag{2.12}

where \( \mathbf{H}_1, \mathbf{H}_2, \) and \( \mathbf{H}_3 \) are of orders \( n \times c \), \( n \times 2c \), and \( n \times c \), respectively. From (2.10) and (2.11) we can write

\[ \lambda = \lambda_{\max}(\mathbf{Q}) + \mathbf{H}_1 \mathbf{H}_2 \mathbf{Y} \]  \tag{2.13}

Consider the following random vector \( \mathbf{Y} \) defined as

\[ \mathbf{Y} = \mathbf{Y} + \left( \lambda_{\max}(\mathbf{Q}) \mathbf{Y} \right) \mathbf{H}_1 \mathbf{Y} \]  \tag{2.14}

where \( \lambda_{\max} \) is the largest eigenvalue of \( \mathbf{Q} \). Note the following (Khuri 1989):

1. \( \text{Cov}(\mathbf{Y}_i, \mathbf{H}_1) = \mathbf{H}_1 \)
2. \( \mathbf{E} \mathbf{H}_1 = \mathbf{H}_1 \mathbf{H}_1 \mathbf{Y} \), where \( \mathbf{Y} \) is the variance-covariance matrix of the vector \( \mathbf{Y} \).

It follows from (2.8) and (2.14) that

\[ \text{Var}(\mathbf{Y}) = \text{Diag}(\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_p) + \lambda_{\max} \mathbf{I} \]  \tag{2.15}

We can then partition \( \mathbf{Y} \) into \( \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p \), such that

\[ \mathbf{Y}_i \sim \text{MVN} \left( \begin{bmatrix} \mathbf{0} \mathbf{I}_1 \mathbf{I}_2 \ldots \mathbf{I}_p \end{bmatrix} \lambda_{\max} \mathbf{I} \right) \]  \tag{2.16}

where \( \mathbf{Y} \) is a vector containing the values of the dependent variable.

The \( \mathbf{Y}_i \)'s are independently distributed as normal random vectors. Furthermore, they are independently distributed from

\[ \mathbf{SS}_1 = \mathbf{Y} \mathbf{H}_2 \mathbf{Y} \]  \tag{2.17}

If we then form the sums of squares \( \mathbf{SS}_i (i = 1, 2, \ldots, p) \) defined by \( \mathbf{SS}_i = \mathbf{Y} \mathbf{H}_2 \mathbf{Y} \), we have that

\[ \mathbf{SS}_i \mathbf{H}_1 (i + \lambda_{\max} \mathbf{Y}^2) \]  \sim \mathbf{X}_{n_1}^2 \tag{2.18}

As shown in Khuri (1989) these quadratic forms provide exact tests for testing most of the variance components in the model. Although, as is true with balanced designs, Satterthwaite's approximation might be needed when testing some of the variance components. This occurs when no \( \mathbf{MS}_i = \mathbf{SS}_i / n_1 \) exists whose expected value is equal to that for the effect being tested.

3. Description of the SAS Macro

The theory outlined above represents the only available methodology for deriving exact tests for random-effects models, hence these techniques are not available in any of the software packages. The following steps outline some of the basic functions used in the SAS software macro developed for deriving these tests. The key computations involved are illustrated using a three-factor model defined by indices \( i, j, \) and \( k \) where \( j \) and \( k \) correspond to two factors that are crossed, but nested within \( i \).

\[ \mathbf{Y}_{ijkl} = \mu + \mathbf{Y}_i + \mathbf{Y}_{ij} + \mathbf{Y}_{ik} + \mathbf{Y}_{ijk} \]  \tag{3.1}

Procedure:

1. The user creates a data set specifying the observations. In the first step of the macro, we obtain:
   (i) the number of levels for each factor in the model;
   (ii) the number of observations in each cell;
   (iii) a vector containing the values of the dependent variable.

The program uses PROC SUMMARY to obtain (i) and (ii).
2. The assumption, \( n > 2\alpha + 1 \), required in (2.11) is verified. If this assumption is not satisfied, the process is terminated and the following error message is printed:

SAMPLE SIZE NOT ADEQUATE FOR EXACT METHOD

3. The macro UNBAL requires only the specification of the model. The model specification is similar to the syntax used in PROC GLM. For instance, cross effects are specified by joining the variables with asterisks such as PLANT*SITE. Nested effects are specified by using a parenthetical field after the variable indicating the variables within which the effect is nested. The macro call involves three parameters (the last one is optional):

\%
UNBAL(response, effects [DATA = dsname]);

where:
response identifies the dependent variable in the model;
effects identifies the categorical variables in the model, the crossed effects, and the nested effects;
dsname identifies the data set (default = _LAST_).

For example,

\%
UNBAL(YDEP, I J(I) K(I) J.K(I), DATA = MYDATA)

4. For each effect in the model, we create two types of sets:

(i) the sets \( S_\ell \) for \( \ell = 1, 2, \ldots, e \) whose elements identify the factors in the effect;

(ii) the sets \( R_\ell \) for \( \ell = 1, 2, \ldots, e \) whose elements identify the factors that are nested in the effect.

A main effect, while not nested, will appear as a set \( R_\ell \). These sets are used in creating the \( P_\ell \)'s in a process defined in Khuri (1982), using components of admissible means. The sets, \( S_\ell \) and \( R_\ell \), facilitate the steps in producing these components.

For model (3.1) the following sets are created:

\[ S_1 = \{I\}, \ S_2 = \{I J\}, \ S_3 = \{I K\}, \ S_4 = \{I J K\} \]
\[ R_1 = \{I\}, \ R_2 = \{J\}, \ R_3 = \{K\}, \ R_4 = \{J K\} \]

5. Create the \( P_\ell \)'s (the matrices associated with the sum of squares for the \( \ell \)th effect for \( \ell = 0, 1, \ldots, n \)). The matrices \( P_\ell \)'s can be expressed in terms of the matrices \( A_\ell \)'s (the \( A_\ell \)'s are given in (2.3)) since

\[ P_1 = \sum \frac{A_\ell}{\lambda_{ij}} h_2 A_j; \]

where \( \lambda_{ij} \) is a known constant with possible values -1, 0, and 1.

The \( P_1 \)'s for model (3.1) are given by:

\[ P_1 = \frac{1}{b_{ij}} A_1 - \frac{1}{b_{ij}} A_0 \]
\[ P_2 = \frac{1}{b_{ij}} A_2 - \frac{1}{b_{ij}} A_1 \]
\[ P_3 = \frac{1}{b_{ij}} A_3 - \frac{1}{b_{ij}} A_1 \]
\[ P_4 = \frac{1}{b_{ij}} A_4 - \frac{1}{b_{ij}} A_2 - \frac{1}{b_{ij}} A_3 + \frac{1}{b_{ij}} A_1 \]

where
\[ A_0 = J_{36}; \quad A_1 = I_3 \otimes J_{12}; \quad A_2 = I_{12} \otimes J_3; \]
\[ A_3 = I_3 \otimes J_4 \otimes I_3; \quad A_4 = I_{36}; \]

6. Create the \( Q_\ell \)'s using the LINDEP operator and the matrix \( Q \) defined in (2.4). Recall that the matrix \( Q_\ell \) is the Gram-Schmidt orthonormalization of the rows of \( P_\ell \). This is accomplished by using the GS function in PROC MATRIX.

Let \( P_1 \) represent the matrix \( P_1 \) and \( Q_1 \) the matrix \( Q_1 \). The result is obtained from:

\[ \text{GS} \ Q_1 \ T \text{LINDEP} \ P_1; \]

7. Create the transformation in (2.14). The two key computations here involve deriving the matrices:

\[ M = (\max (1 - Q^* K Q^*)^{\frac{1}{2}} \text{ and } B_1. \]
MACRO STATEMENTS

G = Q'KQ %STR(');
MAX = MAX(EIGVAL(G));
PSI1 = C * 1;
PSI2 = N - 2*C + 1;
IDEN = I(PSI1);
M = LMAX*IDEN - G;
EIGEN EMLM;
GAM1 = DIAG(SQRT(EM));
MHAL = L*GAM1*L %STR(');

(ii). Create \( \Pi_1 \):
Recall from (2.11) and (2.12) that

\[ R = HAH' = H_1H_1' + H_2H_2'. \]

where the columns of \( R \) are the eigenvectors of the matrix \( R \). Only \( N - c \) of these are independent. We then derive the matrix \( \Pi_1 \) by selecting \( c - 1 \) independent columns of \( H \).

Hence, the first step in obtaining \( \Pi_1 \) is to determine the eigenvectors of \( H \) that are independent and we accomplish this by using the GS function once again. Let \( \Pi_1 \) denote the Gram-Schmidt orthonormalization of \( H \). We then use the statement:

\[ \text{GS PH TIL LINDEP H.} \]

Since \( c \) columns of the matrix \( H \) are linearly independent, \( c \) columns of \( \Pi_1 \) are set to zero. The particular columns that are set to zero correspond to the dependent columns in \( H \). As a result, \( H_2 \) is then obtained by selecting \( c - 1 \) columns in \( H \) that do not correspond to the 0 columns of the matrix \( \Pi_1 \). This is accomplished with the statements:

\[ \text{GS PH TIL LINDEP H; DO } I = 1 \text{ TO } N; \]
\[ \text{IF NCOL(LOC(ROUND(PH(I)))) NE 0 THEN H1 = H1 | H(I); END;} \]
\[ H1 = H1(1:C - 1); \]

Note that the LOC function identifies the rows of \( \Pi_1 \) having nonzero elements. If the row vector created by the function LOC is empty (NCOL = 0), then the \( i \)th column in \( \Pi_1 \) is a column of zeros.

8. Create the mean squares - MS; \( s \):
Having obtained \( w \) in (2.14), we obtain the subvectors \( \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_v \) by using row and column subscripts such as

\[ W_1 = W(1:M_1); \]
\[ W_2 = W(M_1 + 1:M_1 + M_2); \]
\[ \vdots \]

where \( M_1 \) corresponds to the rank of \( \Pi_1 \) or the degrees of freedom for effect 1. We then obtain the MS; \( s \)'s from

\[ SSI = W_1 %STR(') * W_1; \]
\[ MSI = SSI#/M_1; \]

9. Construct the test statistics based on the expected mean squares. We may still find the need to use Satterthwaite's approximation when testing some of the variance components.

10. Output: \( F \)-values, a table of expected mean squares, and a vector containing the test statistics.

FUTURE RESEARCH

Current research will focus on the general mixed model with unequal cell frequencies in the last stage. We have already developed exact tests for the two-way cross classification mixed model (Gallo and Khuri 1988) and the 2-fold nested mixed model (Gallo 1987).

ACKNOWLEDGEMENTS

We would like to thank Dr. John K. Cuddeback, Vice-President of Information Services, Shands Hospital, for providing us with the time to pursue this exciting endeavor.

SAS \( \text{R} \) is the registered trademark of SAS Institute Inc., Cary, NC, USA.

Author Contact
Jose Gallo, Ph.D.
Senior Statistician
Richard MacElderry
Computer Systems Analyst
Division of Information Services
Box J-356, HCA #5
Shands Hospital
Gainesville, FL 32610
(904) 395-4010
REFERENCES


