A Comparison of Two Approaches For Selecting Covariance Structures in The Analysis of Repeated Measurements

by

H.J. Keselman
University of Manitoba

James Algina
University of Florida

Rhonda K. Kowalchuk
University of Manitoba

and

Russell D. Wolfinger
SAS Institute

Abstract

The mixed model approach to the analysis of repeated measurements allows users to model the covariance structure of their data. That is, rather than using a univariate or a multivariate test statistic for analyzing effects, tests that assume a particular form for the covariance structure, the mixed model approach allows the data to determine the appropriate structure. Using the appropriate covariance structure should result in more powerful tests of the repeated measures effects according to advocates of the mixed model approach. SAS' (1996) mixed model program, PROC MIXED, provides users with two information criteria for selecting the `best' covariance structure, Akaike (1974) and Schwarz (1978). Our study compared these log likelihood tests to see how effective they would be for detecting various population covariance structures. In particular, the criteria were compared in unbalanced (across groups) nonspherical repeated measures designs having equal/unequal group sizes and covariance matrices when data were both normally and nonnormally distributed. The results indicate that neither criterion was effective in finding the correct structure. On average, for the 26 investigated distributions, the Akaike criterion only resulted in the correct structure being selected 47 percent of the time while the Schwarz criterion resulted in the correct structure being selected just 35 percent of the time. Not surprisingly, PROC MIXED default F-tests based on either of these selection criteria performed poorly according to results reported by the authors elsewhere.

Key Words: Covariance Structures, Akaike Criterion, Schwarz Criterion, Repeated Measurements

A Comparison of Two Approaches For Selecting Covariance Structures in The Analysis of Repeated Measurements

1. Introduction

The traditional analysis of variance F-tests for repeated measures designs containing between-subjects variables require that the data conform to the multisample sphericity assumption in order to be valid (see Huynh & Feldt, 1970; Keselman & Algina, 1996; Keselman & Keselman, 1993; Rogan, Keselman & Mendoza, 1979; Rouanet & Lepine, 1970). That is, for a J (between-subjects) $\times$ K (within-subjects) design, $F_K = MS_K/MS_{K\times S/J}$ and $F_{JK} = MS_{JK}/MS_{K\times S/J}$ will only be distributed as F variables when, in addition to multivariate normality and independence assumptions, the variances for all possible differences between the levels of the repeated measures variable are equal (i.e., sphericity) and this constant variance holds for each level of the between-subjects grouping variable (i.e., multisample sphericity, see Mendoza, 1980). Equivalently, multisample sphericity can be stated as $\mathbf{C}^T\mathbf{\Sigma}_j\mathbf{C} = \lambda\mathbf{I}_{(K-1)}$, (j = 1, … , J), where $\mathbf{C}$ is an orthonoromalized contrast matrix representing a comparison among the levels of the repeated measures factor, $\mathbf{\Sigma}_j$ is a variance-covariance matrix for K associated with a particular level of treatment j, λ is a scalar > 0, $\mathbf{I}$ is a $K-1 \times K-1$ identity matrix, and ^{T} is the transpose operator.

Since the data obtained in applied settings, particularly in the behavioral sciences, will rarely conform to the multisample sphericity pattern, many authors have recommended that applied researchers adopt a corrected degrees of freedom (df) test, such as Greenhouse and Geisser (1958) or Huynh and Feldt (1976), or a multivariate test statistic, in order to obtain valid tests of repeated measures effects. A corrected df univariate test attempts to circumvent the multisample sphericity assumption by altering the df of the traditional test statistics based on a sample estimate of the unknown sphericity parameter. The multivariate test, on the other hand, does not require sphericity, only covariance homogeneity and multivariate normality. The empirical evidence regarding the validity of these two approaches indicates that if the design is balanced they effectively control the probability of committing a Type I error at the desired level of significance. However, for unbalanced (unequal group sizes) repeated measures designs, neither approach effectively maintains control over Type I errors when covariance heterogeneity exists, a likely outcome with applied data (e.g., see Keselman & Keselman, 1990; Keselman, Lix & Keselman, 1996). That is, current recommendations do not generally hold in unbalanced (across groups) repeated measures designs when covariance matrices are heterogeneous. Unfortunately, according to a survey conducted by Kowalchuk, Lix, and Keselman (1996), repeated measures designs in the applied sciences are typically unbalanced.

One of the newer approaches to the analysis of repeated measurements is based on a mixed model approach (see Jennrich & Schluchter, 1986; Laird & Ware, 1982; Liang & Zeger, 1986; Little, Milliken, Stroup & Wolfinger, 1996). The potential benefit of this approach is that it allows a user to model the covariance structure of the data rather than presuming a certain type of structure as is the case with the traditional univariate and multivariate test statistics. Parsimonously modeling the covariance structure of the data should result in more efficient estimates of the fixed-effects

parameters of the model and consequently more powerful tests of the repeated measures effects. The mixed model approach, and specifically SAS's (SAS, 1996) PROC MIXED, allows users to fit various covariance structures to the data. For example, some of the structures that can be fit with PROC MIXED are: (a) compound symmetric (CS), (b) unstructured (UN), (c) spherical/Huynh and Feldt (1970) (HF), (d) first order autoregressive (AR), and (e) random coefficients (RC). The CS structure is assumed by the traditional univariate F-tests in SASs GLM program (SAS Institute, 1990), while the UN structure is assumed by GLMs multivariate tests of the repeated measures effects. AR and RC structures more appropriately reflect that measurement occassions that are closer in time are more highly correlated than those farther apart in time. In addition, PROC MIXED allows users to specify, separately and jointly, between-subjects and within-subjects heterogeneity. After a covariance structure is selected, SAS computes, by default, F-tests which are Wald-type statistics which are asymptotically valid and whose sampling distribution is approximated by an F in small samples (see McLean & Sanders, 1988; Wolfinger, 1993).

It is suggested that users first determine the appropriate covariance structure prior to conducting tests of significance for the repeated measures effects (see Little et al., 1996; Wolfinger, 1993; 1996). Specifically, a covariance structure can be selected by comparing the Akaike Information Criterion (AIC) (Akaike, 1974) and/or Schwarz Bayesian Criterion (BIC) (Schwarz, 1978) values for various potential covariance structures (see Bozdogan, 1987; Little et al; Wolfinger, 1993; 1996). According to the authors of SAS manuals these two criteria are likely to result in basically equivalent results. However, to date, we know of no reported study which has examined this proposition. Thus, the purpose of our investigation was to compare the two criteria.

2. Methods

The simplest of the higher-order repeated measures designs involves a single between-subjects factor and a single within-subjects factor, in which subjects ($i = 1, \ldots, n_j$, $\Sigma n_j = N$) are selected randomly for each level of the between-subjects factor ($j = 1, \ldots, J$) and observed and measured under all levels of the within-subjects factor ($k = 1, \ldots, K$). In this design, the repeated measures data are modeled by assuming that the observations Y_{ijk} are normal, independent and identically distributed within each level j, with common mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$.

2.1 Test Statistics

The AIC and BIC information criteria, in larger-is-better form, can be specified as

$$\mathrm{AIC}_R = l_R(\widehat{\theta}) - q$$

and

$$\mathrm{BIC}_R = l_R = l_R(\widehat{\theta}) - \tfrac{q}{2}\log(n - p),$$

where $\widehat{\theta}$ is the restricted /residual maximum likelihood estimate of the unknown variance-covariance parameters θ, q is the number of elements in θ, and for the model given previously $p = JK$ and $n = NK$ (See Wolfinger, 1996).[1]

2.2 Study Variables

The conditions of our study were those used by Keselman, Algina, Kowalchuk and Wolfinger (1997) in their study which compared various approaches, including the mixed model approach, to the analysis of repeated measurements. Thus, the conditions (e.g., covariance heterogeneity, nonnormality, unequal group sizes) selected were chosen primarily for their known effects on tests for mean equality. Nonetheless, we believe that the conditions they manipulated would be relevant as well in our investigation, when examining the log likelihood tests, since they were chosen to mirror conditions likely to be encountered by researchers working in applied settings.

The two criteria (AIC and BIC) for selecting a covariance structure prior to computing tests for testing repeated measures effects were examined for balanced and unbalanced designs containing one between-subjects and one within-subjects factor; there were three and four levels of these factors, respectively. Selected combinations of six factors were investigated.

Six covariance structures investigated were: (a) UN, (b) ARH, and (c) RCH, (d) UN_j, (e) ARH_j, and (f) RCH_j, where H in the structure designation refers to within-subject heterogeneity and the subscript $_j$ denotes between groups heterogeneity (i.e., allowing for heterogeneity of a structure across groups) (see Little et al., 1996; SAS, 1995; Wolfinger, 1993; 1966). The authors will provide upon request an enumeration of the element values used in the simulation study.

For each of the preceding structures, equal (excluding RCH) as well as unequal between-subjects covariance matrices were investigated. When unequal, the elements of the matrices were in the ratio of 1:3:5.

Based on the belief that applied researchers work with data that is characterized by both within and between heterogeneity, eleven covariance structures were fit with PROC MIXED for the Akaike (1974) and Schwarz (1978) criteria. These structures were: (a) CS, (b) UN, (c) AR, (d) HF, (e) CSH, (f) ARH, (g) RCH, (h) UN_j, (i) HF_j, (j) ARH_j, and (k) RCH_j. Thus, we allowed PROC MIXED to select from among homogeneous, within heterogeneous, and within and between heterogeneous structures.

The criteria were investigated when the number of observations in each group were equal or unequal. Based on the findings reported by Keselman et al. (1993) and Wright (1995) we considered, like Keselman et al. (1997), a number of cases of total sample size: $N = 30$, $N = 45$, and $N = 60$. For each value of N, both a moderate and substantial degree of group size inequality were typically investigated. The moderately unbalanced group sizes had a coefficient of sample size variation (C) equal to $\simeq .16$, while for the more disparate cases $C \simeq .33$, where C is defined as $(\Sigma_j(n_j - \overline{n})^2/J)^{\frac{1}{2}}/\overline{n}$, and $\overline{n}$ is the average group size. The $C \simeq .16$ and $C \simeq .33$ unequal group sizes cases were respectively equal to: (a) 8, 10, 12 and 6, 10, 14 ($N = 30$), (b) 12, 15, 18 and 9, 15, 21 ($N = 45$), and (c) 16, 20, 24 and 12, 20, 28 ($N = 60$).

Positive and/or negative pairings of these unequal group sizes and unequal covariance matrices were investigated. A positive pairing referred to the case in which the largest n_j was associated with the covariance matrix containing the largest element

values; a negative pairing referred to the case in which the largest n_j was associated with the covariance matrix with the smallest element values. In short, for each value of N, four pairings of unequal covariance matrices and unequal group sizes were investigated: moderately and very unequal n_js which were both positively and negatively paired with the unequal $\mathbf{\Sigma}_j$s.

The sphericity index was set at $\epsilon = 0.75$. The 0.75 value characterizes data found in educational and behavioral science research (Huynh & Feldt, 1976). When $\epsilon = 1.0$, sphericity is satisfied and for the J × K design the lower bound of $\epsilon = 1/(K - 1)$.

The criteria were compared when the simulated data were obtained from multivariate normal or multivariate nonnormal distributions. The nonnormal distribution was a multivariate lognormal distribution with marginal distributions based on $Y_{ijk} = \exp(X_{ij})$ $(i = 1, \ldots, n_j)$ where X_{ijk} is distributed as N(0, .25); this distribution has skewness (γ_1) and kurtosis (γ_2) values of 1.75 and 5.90, respectively. The algorithms for generating the two investigated distributions can be found in Keselman et al (1993) and Algina and Oshima (1995).

An enumeration of the conditions investigated can be found in Table 1. One thousand replications of each condition were performed.

Insert Table 1 About Here
---

3. Results

In the tables that follow we present results for a subset of the selected combinations investigated; the subset adequately demonstrates differences that exist between the criteria.

Tables 2 and 3 contain percentages, for each condition investigated, of the three most frequently selected covariance structures by the AIC and BIC criteria, respectively. Shaded cells of the table indicate the true covariance structure of the data.

Insert Tables 2 and 3 About Here

The results in Table 2 indicate that, for the AIC criterion,: (a) on average, across the 26 investigated conditions, the correct covariance struture is selected 47 percent of the time, (b) when the true structure is UN/UN_j-Normal (A-F), another structure (ARH_j) is selected more often than the correct structure, (c) the correct structure is selected more frequently than any of the other structures when the true structure is ARH_j-Normal (H-L), RCH_j-Normal (M-Q), UN_j-Lognormal data (R-S), and ARH_j-Lognormal data (T-U) and (d) the correct structure (RCH_j) is selected approximately as frequently as an incorrect structure (UN_j) for Lognormal data (V-Z).

The picture is quite different when applying the Schwarz criterion. In particular, (a) in 14 of the 26 conditions investigated the correct covariance structure was never selected, (b) averaged over the cases in which the correct covariance structure was selected some percentage of the time, the percent of correct selections was 35, and (c) for

all structures, except RCH_j-Lognormal (V-Z) and distribution Q, an incorrect structure was selected more frequently than the correct structure.

4. Discussion

The newest approach to the analysis of repeated measurements is a mixed model analysis. Advocates of this approach suggest that it provides the `best' approach to the analysis of repeated measurements since it can, among other considerations, handle missing data and allow users to model the covariance structure of the data. The first of these advantages is typically not a pertinent issue to those involved in controlled experiments since data in these contexts is rarely missing. The second consideration, however, could be most relevant to experimenters since, according to the developers of mixed model analyses, modeling the correct covariance structure of the data should result in more powerful tests of the fixed-effects parameters. That is, users should obtain more powerful tests by using test statistics that more accurately model the correct covariance structure rather than adopting the usual univariate or multivariate tests which presume specific types of covariance structures for the data.

The mixed model program in SAS (1996) allows users to model many potentially applicable covariance structures. Additionally, the program allows even greater flexibility to the user by allowing him/her to model covariance structures that have within-subjects and/or between-subjects heterogeneity. In order to select an appropriate structure for ones data, PROC MIXED users can use either an Akaike (1974) or Schwarz (1978) information criteria. These log likelihood tests, it is believed, should provide equivalent results. Nonetheless, within the context of unbalanced (across groups) nonspherical heterogeneous repeated measures designs, comparisons of the crirteria have not to date been made. Accordingly, we compared these criteria for various between- by within-subjects repeated measures designs in which we varied the true covariance structure to the data, the distributional form of the data, as well as group size and covariance balance/imbalance.

Our data indicate that neither approach uniformly selected the correct covariance structure. Indeed, for most of the investigated structures investigated, the Akaike (1974), and in particular the Schwarz (1978) criteria, more frequently picked the wrong covariance structure. Averaging over the conditions in which a correct structure was selected some percentage of the time, the correct structure was picked 47 percent of the time with the Akaike criterion and only 35 percent of the time with the Schwarz criterion. Thus, though the mixed model approach allows users to model the covariance structure, two popular criteria for selecting the `best' structure perform poorly. Not surprisingly, Keselman et al. (1997) found that the default F-tests that PROC MIXED computes based on either of these two criteria were prone to inflated rates of Type I error. Accordingly, potential presumed power benefits must be discounted when the procedure is prone to excessive rates of Type I error.

The authors are currently examining whether PROC MIXED tests of repeated measures effects can be improved by using its Satterthwaite test option. In the meantime we continue to recommend that for the analysis of repeated measures effects users adopt the nonpooled multivariate Welch-type statistic presented by Keselman, Carriere and Lix (1993). For most unbalanced nonspherical heterogeneous repeated measures designs it typically will be robust even when data is nonnormal (see Keselman et al., 1993). Lix

and Keselman (1995) show how to apply this statistic to test omnibus and subeffect tests in most independent and correlated groups designs and present a SAS/IML (SAS, 1989) program to obtain numerical results.

FOOTNOTES

1. Release 7.01 of PROC MIXED will compute the Schwarz (1978) criteria with a less stringent penalty. Specifically, based on Carlin and Louis (1996), n will be equated with the number of subjects rather than the number of observations as is currently the case in PROC MIXED releases other than 7.01.

REFERENCES

Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transaction on Automatic Control*, AC-19, 716-723.

Algina, J. (in press), "Generalization of Improved General Approximation tests to split-plot designs with multiple between-subjects factors and/or multiple within-subjects factors," *British Journal of Mathematical and Statistical Psychology*.

Algina, J. (1994), "Some alternative approximate tests for a split plot design," *Multivariate Behavioral Research*, 29, 365-384.

Bozdogan, H., (1987), "Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions," *Psychometrika*, 52, 345-370.

Carlin, B.P., and Louis, T.A. (1996), "*Bayes and empirical bayes methods for data analysis*," Chapman and Hall.

Greenhouse, S.W. and Geisser, S. (1959), "On methods in the analysis of profile data," *Psychometrika*, 24, 95-112.

Hotelling, H. (1931), "The generalization of Student's ratio," *Annals of Mathematical Statistics*, 2, 360-378.

Huynh, H.S. and Feldt, L. (1970), "Conditions under which mean square ratios in repeated measurements designs have exact F distributions," *Journal of the American Statistical Association*, 65, 1582-1589.

Huynh, H. and Feldt, L.S. (1976), "Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs," *Journal of Educational Statistics*, 1, 69-82.

Jennrich, R.I., and Schluchter, M.D., (1986), "Unbalanced repeated-measures models with structured covariance matrices," *Biometrics*, 42, 805-820.

Keselman, H.J., Algina, J., Kowalchuk, R.K., and Wolfinger, R.D. (1997), "A comparison of recent approaches to the analysis of repeated measurements," Unpublished manuscript.

Keselman, H.J., Carriere, K.C., and Lix, L.M. (1993), "Testing repeated measures hypotheses when covariance matrices are heterogeneous," *Journal of Educational Statistics*, 18, 305-319.

Kowalchuk, R.K., Lix, L.M., and Keselman, H.J. (1996), "The analysis of repeated measures designs," paper presented at the Annual Meeting of the Psychometric Society, 1996, Banff, Alberta.

Laird, N., and Ware, J.H., (1982), "Random-effects models for longitudinal data," *Biometrics*, 38, 963-974.

Liang, K.Y., and Zeger, S.L., (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13-22.

Little, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D., (1996), *SAS system for mixed models*, Cary, NC: SAS Institute.

Lix, L.M., and Keselman, H.J. (1995), "Approximate degrees of freedom tests: A unified perspective on testing for mean equality," *Psychological Bulletin*, 117, 547-560.

McLean, R.A., and Sanders, W.L. (1988). "Approximating degrees of freedom for standard errors in mixed linear models," *Proceedings of the Statistical Computing Section, American Statistical Association*, New Orleans, pp. 50-59.

Rouanet, H. and Lepine, D. (1970), "Comparison between treatments in a repeated measures design: ANOVA and multivariate methods," *British Journal of Mathematical and Statistical Psychology*, 23, 147-163.

SAS Institute. (1989), *SAS/IML sowtware: Usage and reference, Version 6*. Cary, NC: Author.

SAS Institute. (1990), *SAS/STAT User's Guide Vol. 2, GLM-VARCOMP, Version 6, 4th Ed*. Cary, NC: Author.

SAS Institute. (1995), *Introduction to the MIXED procedure: Course Notes*, Cary, NC: Author.

SAS Institute. (1996), *SAS/STAT Software: Changes and Enhancements through Release 6.11*, Cary, NC: Author.

Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461-464.

Wolfinger, R., (1993), "Covariance structure selection in general mixed models," *Communication in Statistics – Simulation*, 22, 1079-1106.

Wolfinger, R.D., (1996), "Heterogeneous variance-covariance structures for repeated measurements," *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 205-230.

Wright, S.P., (1995), "Adjusted F tests for repeated measures with the MIXED procedure," *Proceedings of the Twentieth Annual SAS Users Group International Conference*, Cary, NC.

Wright, S.P., and Wolfinger, R.D., (1996), "Repeated measures analysis using mixed models: Some simulation results," *Paper presented at the Conference on Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*, Nantucket, MA (October).

Author's Note

This research was supported by a Social Sciences and Humanities Research Council grant (#410-95-0006) to the first author.