Journal of Mathematics and Statistics 10 (3): 309-315, 2014
ISSN: 1549-3644

doi:10.3844/jmssp.2014.309.315 Published Online 10 (3) 2014 (http://www.thescipub.com/jmss.toc)

SELECTING THE COVARIANCE STRUCTURE IN MIXED MODEL USING STATISTICAL METHODS CALIBRATION

Ali Hussein AL-Marshadi

Department of Statistics, King Abdul Aziz University, Jeddah, Saudi Arabia

Received 2014-04-29; Revised 2014-05-24; Accepted 2014-07-08

ABSTRACT

In this article the analysis of experiment of repeated measures design is considered which is used often in different fields of studies. In order to analyze the experiment of repeated measures design efficiently we need to select the suitable covariance structure which required a lot of efforts. In the current paper an approach is used to guide the selection of the covariance structure for the analysis of repeated measures design with high rate of success. Five well known model selection criteria are used in the approach. Simulation study is used to evaluate the approach in terms of its ability to select the right covariance structure. The evaluation of the approach was in terms of its percentage of times that it identifies the right covariance structure. Overall, the simulation study showed excellent performance for the approach in all the study cases. The main result of our article is that we recommend considering the approach as a standard way to select the right covariance structure.

Keywords: Repeated Measures Design, Information Criteria, Bootstrap Procedure, Hierarchical Clustering Methods, Single Linkage Distance Measure, Kenward-Roger Method, Restricted Maximum Likelihood (REML)

1. INTRODUCTION

The correct analysis of a study according to the design of experiment used is very important factor to the success of any study. An inaccurate analyzed of a study can produce misleading results for that study. Repeated measures experimental designs require special attention, since in practice the observations within each subject are more likely to be correlated with different covariance structures that makes their analysis different from other factorial experiments (Bellavance *et al*., 1996; Gill, 1992; McCulloch, 2003). Considering the right covariance structure for the observations within each subject is an important aspect of the analysis of repeated measures experiment; this is where the dependency due to the repeated measures is taken into account.

The mixed procedure of the SAS System is used for analyzing data of repeated measures experiment since it has the capability of fitting the data with different covariance structure according to linear mixed model setup (Littell *et al*., 1999). There was a lot of attention in the earlier history of the linear mixed model on adequately modeling the covariance structure (Chi and Reinsel, 1989; Diggle, 1988; Goldstein *et al*., 1994; Keselman, *et al*., 1998; 1999a; Núñez-Antón and Zimmerman, 2000). Therefore the first step need to be considered in the statistical analysis of data of repeated measures experiment is deciding what the suitable covariance structure of the data is. Researchers often use the information criteria such as AIC, (Akaike, 1974), BIC, (Schwarz, 1978), CAIC, (Bozdogan, 1987), HQIC, (Hannan and Quinn, 1979) and AICC, (Hurvich and Tsai, 1989), for deciding what the correct covariance structure is according on the observed data (Keselman *et al*., 1999b; Littell *et al*., 2000; Singer, 1998). Many studies have investigated the performance of those information criteria in selection of the covariance structure (Yanosky, 2007;

Ferron *et al*., 2002; Gomez *et al*., 2005; Guerin and Stroup, 2000; Keselman *et al*., 1999b). Unfortunately, those criteria do not always select the correct covariance structure and thus possible impacted of misspecification of the covariance structure on statistical properties of the inferences must be taken to account (AL-Marshadi, 2008; Yanosky, 2007; Ferron *et al*., 2002; Gomez *et al*., 2005; Guerin and Stroup, 2000; Keselman *et al*., 1999a). Ferron *et al*. (2002) found that the AIC on average select the correct covariance structure about 79% of the time and the SBC select the correct covariance structure less frequently, on average 66% of the time. In contrast, (Keselman *et al*., 1998) found that the AIC and SBC were only success in select the correct covariance structure 47 and 35% of the time respectively. Resent a Monte Carlo simulation study investigated the misspecification impact of the covariance matrix in the linear mixed model (Brandon, 2013).

Our research objective is evaluating the approach in selecting the right covariance structure, where the evaluation of the approach was in terms of its ability to identify the right covariance structure.

2. METHODOLOGY

The design of the simulated experiment is quite similar to the setup used in (AL-Marshadi, 2008) which is described below.

The treatments were arranged in a basic form of repeated measures design which consists of a completely randomized experimental design with data collected in a sequence of equally spaced points in time. The design of the simulated experiment is consists of:

t = 3 treatments
r = 7 or 10 subjects per treatment level and
a = 7 repeated measures within each treatment level

In mixed procedure, five model selection criteria available, which can be used to select an appropriate covariance structure. The five model selection criteria are:

- Akaike Information Criterion (AIC), (Akaike, 1974)
- Schwarz Bayesian Information Criterion (BIC), (Schwarz, 1978)
- Bozdogan Corrected Akaike Information Criterion (CAIC), (Bozdogan, 1987)
- Hannan and Quinn Information Citerion (HQIC), (Hannan and Quinn, 1979)
- Hurvich and Tsai the corrected Akaike Information Criterion (AICC), (Hurvich and Tsai, 1989)

In this study the previous five information criterions were used with the approach and the approach were evaluated in terms of its ability to identify the right covariance structure.

The algorithm of the approach involves using the bootstrap technique (Efron, 1983; 1986) and hierarchical clustering methods with single linkage distance measure approach (Khattree and Naik, 2000) as tools to calibrate with the five information criterion in selecting the right covariance structure. The idea of using the bootstrap to improve the performance of a model selection rule was introduced by (Efron, 1983; 1986) and is extensively discussed by (Efron and Tibshirani, 1993).

In the context of the mixed model, the algorithm for using the approach can be outlined as follows.

Let the vector y_{ij} is defined as follows: $\boldsymbol{y_{ij}} = \left[y_{ij_1} \quad y_{ij2} \quad ... \quad y_{ija} \right]` \approx MVN(\boldsymbol{u_i}, \Sigma_{all} = \mathbf{J}\sigma^2_{subj} + \Sigma)$ where, μ_i is the vector of means at *kth* time for the *ith* treatment, i.e., $\boldsymbol{\mu_i} = [\mu_{i1} \quad \mu_{i2} \quad ... \quad \mu_{ia}]$, $\mu_{ik} = \alpha_i + \tau_k + (\alpha\tau)_{ik}$ and i = 1, 2,….,t; j = 1, 2,….,r; k = 1, 2,….,a., (AL-Marshadi, 2008):

1. Generate the bootstrap sample on case-by-case using the observed data (original sample) (i.e., based on resampling from ($\mathbf{y_{i1}}$ $\mathbf{y_{i2}}$,…., $\mathbf{y_{ir}}$) for each of the *ith* group independently from the others). The bootstrap sample size is taken to be the same as the size of the observed sample (i.e., r). The properties of the bootstrap when the bootstrap sample size is equal to the original sample size are discussed by (Efron and Tibshirani, 1993)
2. Fit the mixed model with the candidate covariance structures, which we would like to select the right covariance structure from them, to the bootstrap data, thereby obtaining the bootstrap $AIC^*, BIC^*, CAIC^*, HQIC^*, AICC^*$ for the model with the candidate covariance structures
3. Repeat steps (1) and (2) (W) times
4. Researchers often use the previous collection of information criteria in the selection of the right model such as selecting the model with the smallest value of the information criteria (Keselman *et al*., 1999a; Littell *et al*., 2000; Singer, 1998). We will follow

different rule in our algorithm. Bootstrapping of the collected data given us the advantage that for each model and each information criteria we have (W) replication values (from step (1to 3). To make use of this advantage, we propose using the average of each information criteria for each model separately in the algorithm as a random vector that follows 5-dimensional multivariate normal distribution:

$$\left[\overline{AIC} \quad \overline{BIC} \quad \overline{CAIC} \quad \overline{HQIC} \quad \overline{AICC}\right]_{Model-i}$$

To justify that in short, let us consider each model separately and then each average of information criteria approximately follows normal distribution according to central limit theorem. Therefore, we can consider the averages of the information criteria of each model as a random vector that follows 5-dimensional multivariate normal distribution. In this stage Clustering method will play the main rule in our algorithm by clustering the models of candidate covariance structures to two clusters according to the five correlated variables (the averages of the five information criteria). One of the two clusters will be called the cluster of the best set of models of covariance structures. The cluster of the best set of models will be determined according to the cluster that contains the model of general covariance structure UN (Unstructured covariance structure). Then the best model of covariance structure will be the model of simplest covariance structure in the cluster of the best set of models of covariance structures. We refer to our approach as the Approach of Collaboration of Statistical Methods in Selecting the Correct Covariance Structure (ACSMSCCS).

3. THE SIMULATION STUDY

A simulation study of mixed model analysis of repeated measures data was conducted to evaluate ACSMSCCS approach in terms of its percentage of times that it identifies the right covariance structure. Kenward and Roger (1997) was considered for computing the denominator degrees of freedom for the tests of fixed effects from all the analyses in this study, where data are simulated under the null hypothesis. Also, the percentage of times that REML failing to converge with normal situation was reported, when the PROC MIXED procedure used REML without any interfering.

Correlated multivariate normal data were generated according to the described experiment (AL-Marshadi, 2008). There were 12 scenarios to generate data involving six covariance structures and two different sample sizes (r = 7 and 10 subjects per treatment). For each scenario, 5000 datasets were simulated. SAS PROC IML code was written to generate the datasets according to the described design (AL-Marshadi, 2008). The algorithm of ACSMSCCS approach was applied to each one of the 5000 generated data sets. The Percentage of times that the ACSMSCCS approach selects the right covariance structure was reported.

Six common covariance matrix structures were used to simulate correlated error models for the simulated experiment. The six settings of the common covariance matrix are given in **Table 1** which can be categorized to six common covariance structures. The first one, (Setting No. 1) represents Compound Symmetry (CS) covariance structures. The second one, (Setting No. 2) represents first-order Autoregressive (AR) (1) covariance structure. The third one, (Setting No. 3) represents Toeplitz (TOEP) covariance structure. The fourth one, (Setting No. 4) represents Heterogeneous Compound Symmetry (HCS) covariance structure. The fifth one, (Setting No. 5) Heterogeneous first-order Autoregressive (ARH) (1) covariance structure. The sixth one, (Setting No. 6) represents Unstructured (UN) covariance structure.

4. RESULTS

Table 2 summarizes results of the percentage of times that the ACSMSCCS approach selects the right covariance structure from the six Covariance structures, when W = 10, r = 7. **Table 3** summarizes results of the percentage of times that the ACSMSCCS approach selects the right covariance structure from the six Covariance structures, when W = 10, r = 10. Also, the comparison of the results in **Table 2** and **3** may suggest that the performance of the approach improved with increasing of sample size.

Finally, **Table 4** showed the percentage of times that the PROC MIXED procedure failing to converge when the PROC MIXED procedure used REML without any interfering for all the investigated settings of covariance matrix and W = 10 and r = 7. **Table 5** showed the percentage of times that the PROC MIXED procedure failing to converge when the PROC MIXED procedure used REML without any interfering for all the investigated settings of covariance matrix and W = 10 and r = 10. In general, the comparison of the results in **Table 4** and **5** may suggest that the convergence problem could be overcome with the increasing the of sample size.

Table 1. The 6 settings of the covariance matrix structures used in the simulations

Setting No.	Covariance matrix
1	$\begin{bmatrix} 16 & 12.8 & 12.8 & 12.8 & 12.8 & 12.8 & 12.8 \\ 12.8 & 16 & 12.8 & 12.8 & 12.8 & 12.8 & 12.8 \\ 12.8 & 12.8 & 16 & 12.8 & 12.8 & 12.8 & 12.8 \\ 12.8 & 12.8 & 12.8 & 16 & 12.8 & 12.8 & 12.8 \\ 12.8 & 12.8 & 12.8 & 12.8 & 16 & 12.8 & 12.8 \\ 12.8 & 12.8 & 12.8 & 12.8 & 12.8 & 16 & 12.8 \\ 12.8 & 12.8 & 12.8 & 12.8 & 12.8 & 12.8 & 16 \end{bmatrix}$
2	$\begin{bmatrix} 16 & 14.4 & 12.96 & 11.664 & 10.4976 & 9.44784 & 8.503056 \\ 14.4 & 16 & 14.4 & 12.96 & 11.664 & 10.4976 & 9.44784 \\ 12.96 & 14.4 & 16 & 14.4 & 12.96 & 11.664 & 10.4976 \\ 11.664 & 12.96 & 14.4 & 16 & 14.4 & 12.96 & 11.664 \\ 10.4976 & 11.664 & 12.96 & 14.4 & 16 & 14.4 & 12.96 \\ 9.44784 & 10.4976 & 11.664 & 12.96 & 14.4 & 16 & 14.4 \\ 8.503056 & 9.44784 & 10.4976 & 11.664 & 12.96 & 14.4 & 16 \end{bmatrix}$
3	$\begin{bmatrix} 16 & 1.6 & 8 & 6.4 & 4.8 & 3.2 & 11.2 \\ 1.6 & 16 & 1.6 & 8 & 6.4 & 4.8 & 3.2 \\ 8 & 1.6 & 16 & 1.6 & 8 & 6.4 & 4.8 \\ 6.4 & 8 & 1.6 & 16 & 1.6 & 8 & 6.4 \\ 4.8 & 6.4 & 8 & 1.6 & 16 & 1.6 & 8 \\ 3.2 & 4.8 & 6.4 & 8 & 1.6 & 16 & 1.6 \\ 11.2 & 3.2 & 4.8 & 6.4 & 8 & 1.6 & 16 \end{bmatrix}$
4	$\begin{bmatrix} 4 & 4.8 & 6.4 & 8 & 9.6 & 11.2 & 12.8 \\ 4.8 & 9 & 9.6 & 12 & 14.4 & 16.8 & 19.2 \\ 6.4 & 9.6 & 16 & 16 & 19.2 & 22.4 & 25.6 \\ 8 & 12 & 16 & 25 & 24 & 28 & 32 \\ 9.6 & 14.4 & 19.2 & 24 & 36 & 33.6 & 38.4 \\ 11.2 & 16.8 & 22.4 & 28 & 33.6 & 49 & 44.8 \\ 12.8 & 19.2 & 25.6 & 32 & 38.4 & 44.8 & 64 \end{bmatrix}$
5	$\begin{bmatrix} 4 & 4.8 & 5.12 & 5.12 & 4.9152 & 4.58752 & 4.194304 \\ 4.8 & 9 & 9.6 & 9.6 & 9.216 & 8.60160 & 7.86432 \\ 5.12 & 9.6 & 16 & 16 & 15.36 & 14.336 & 13.1072 \\ 5.12 & 9.6 & 16 & 25 & 24 & 22.4 & 20.48 \\ 4.9152 & 9.216 & 15.36 & 24 & 36 & 33.6 & 30.72 \\ 4.58752 & 8.6016 & 14.336 & 22.4 & 33.6 & 49 & 44.8 \\ 4.194304 & 7.86432 & 13.1072 & 20.48 & 30.72 & 44.8 & 64 \end{bmatrix}$
6	$\begin{bmatrix} 4 & 2.4 & 4.8 & 8 & 8.4 & 7 & 4.96 \\ 2.4 & 9 & 2.4 & 1.5 & 2.7 & 7.35 & 10.8 \\ 4.8 & 2.4 & 16 & 3.4 & 10.08 & 15.4 & 6.48 \\ 8 & 1.5 & 3.4 & 25 & 18.9 & 16.45 & 9.2 \\ 8.4 & 2.7 & 10.08 & 18.9 & 36 & 4.62 & 22.56 \\ 7 & 7.35 & 15.4 & 16.45 & 4.62 & 49 & 16.24 \\ 4.96 & 10.8 & 6.48 & 9.2 & 22.56 & 16.24 & 64 \end{bmatrix}$

Table 2. The percentage of times that the ACSMSCCS approach selects the true covariance structures from the possible Covariance structures when r = 7 and W = 10

The correct model	The cluster of the best set of covariance structures	The percent of success (%)
CS	CS, CSH, TOEP, TOEPH, UN	98.16
AR (1)	AR (1), ARH (1), TOEP, TOEPH, UN	99.54
TOEP	TOEP, TOEPH, UN	97.52
CSH	CSH, ARH (1), TOEPH, UN	98.34
ARH (1)	ARH (1), TOEPH, UN	96.56
UN	UN	97.32
Over all the percent of success		97.91

Table 3. The percentage of times that the ACSMSCCS approach selects the true covariance structures from the possible Covariance structures when r = 10 and W = 10

The correct model	The cluster of the best set of covariance structures	The percent of success (%)
CS	CS, CSH, TOEP, TOEPH, UN	97.54
AR (1)	AR (1), ARH (1), TOEP, TOEPH, UN	99.74
TOEP	TOEP, TOEPH, UN	97.12
CSH	CSH, ARH (1), TOEPH, UN	99.14
ARH (1)	ARH (1), TOEPH, UN	97.00
UN	UN	97.78
Over all the percent of success		98.05

Table 4: The Percentage of times that the PROC MIXED procedure failing to converge when the PROC MIXED procedure used REML without any interfering for all the investigated settings of covariance matrix and W = 10 and r = 7

	The right covariance structure						
The fitted structure	AR(1) (%)	ARH(1) (%)	CS (%)	CSH (%)	TOEP (%)	TOEPH (%)	UN (%)
CS	0	0	0	0.00	0	0.00	0.04
AR (1)	0	0	0	0.00	0	0.00	0.08
TOEP	0	0	0	0.00	0	0.02	0.04
CSH	0	0	0	0.00	0	0.00	0.22
ARH (1)	0	0	0	0.00	0	0.02	0.20
UN	0	0	0	0.04	0	0.02	9.18

Table 5. The Percentage of times that the PROC MIXED procedure failing to converge when the PROC MIXED procedure used REML without any interfering for all the investigated settings of covariance matrix and W = 10 and r = 10

The fitted structure	The right covariance structure						
	AR (1) (%)	ARH (1) (%)	CS (%)	CSH (%)	TOEP (%)	TOEPH (%)	UN (%)
CS	0	0	0	0	0	0.02	0.00
AR (1)	0	0	0	0	0	0.00	0.00
TOEP	0	0	0	0	0	0.00	0.00
CSH	0	0	0	0	0	0.00	0.02
ARH (1)	0	0	0	0	0	0.00	0.00
UN	0	0	0	0	0	0.00	12.80

5. CONCLUSION

In our simulation, we considered repeated measure design, looking at the performance of the ACSMSCCS approach for selecting the right covariance structure with different settings of the covariance structures. Overall, the ACSMSCCS approach provided excellent tool to select the right covariance structure under the investigated covariance structures. In future studies, it would be interesting to investigate the performance of the approach with other experimental designs such as repeated repeated measure design where there are two levels of repeated measures.

In addition, there is a need to consider more covariance structures and clustering algorithm in the future studies.

6. REFERENCES

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automatic Cont., 19: 716-723. DOI: 10.1109/TAC.1974.1100705

AL-Marshadi, A.H., 2008. The impact of restricted our analysis of repeated measures design to the two stander covariance structures with and without missing data. Aust. J. Basic Applied. Sci., 2: 1228-1238.

Bellavance, F., S. Tardif and M. Stephens, 1996. Tests for the analysis of cross-over designs with correlated errors. Biometrics, 52: 607-612.

Bozdogan, H., 1987. Model selection and Akaike's information criterion (aic): The general theory and its analytical extensions, Psychometrika, 52: 345-370. DOI: 10.1007/BF02294361

Brandon, L., 2013. Misspecification of the covariance matrix in the linear mixed model: A Monte Carlo simulation. University of Minnesota, MN, USA, 2013: 159-159.

Chi, E.M. and G.C. Reinsel, 1989. Models for longitudinal data with random Effects and AR(1) errors. J. Am. Statis. Assoc., 84: 452-459. DOI:10.1080/01621459.1989.10478790

Diggle, P.J., 1988. An approach to the analysis of repeated measurements. Biometrics, 44: 959-971. PMID: 3233259

Efron, B. and R.J. Tibshirani, 1993. An Introduction to the Bootstrap. 2nd Edn., CRC Press, ISBN-10: 0412042312, pp: 456.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Statis. Assoc., 78: 316-331.

Efron, B., 1986. How biased is the apparent error rate of a prediction rule?. J. Am. Statis. Assoc., 81: 461-470. DOI: 10.1080/01621459.1986.10478291

Ferron, J., R. Dailey, Q. Yi, 2002. Effects of misspecifying the first-level error structure in two-level models of change. Multivariate Behav. Res., 37: 379-403. DOI: 10.1207/S15327906MBR3703_4

Gill, P.S., 1992. Balanced change-over designs for auto correlated observations. Aus. J. Statis., 34: 415-420. DOI: 10.1111/j.1467-842X.1992.tb01057.x

Goldstein, H., M. Healy and J. Rasbash, 1994. Multilevel time series models with applications to repeated measures data. Statis. Med., 13: 1643-1655. DOI: 10.1002/sim.4780131605

Gomez, E.V., G.B. Schaalje and G.W. Fellingham, 2005. Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. Commun. Stat. Simul. Comput., 34: 377-392. DOI: 10.1081/SAC-200055719

Guerin, L. and W.W. Stroup, 2000. A simulation study to evaluate PROC MIXED Analysis of repeated measures data. Proceedings of the Conference on Applied Statistics in Agriculture, Manhattan, (SAM' 00), KS, Kansas State University, pp: 170-203.

Hannan, E.J. and B.G. Quinn, 1979. The Determination of the order of an autoregression. J. Royal Stat. Soc. Ser., B, 41: 190-195.

Hurvich, C.M. and C.L. Tsai, 1989. Regression and time series model selection in small samples. Biometrika, 76: 297-307. DOI: 10.1093/biomet/76.2.297

Kenward, M.G. and J.H. Rogers, 1997. Small sample inference for fixed effect from restricted maximum likelihood. Biometrics, 53: 983-997.

Keselman, H.j., J. Algina, R.K. Kowalchuk and R.D. Wolfinger, 1998. A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. Commun. Stat. Simulat. Comput., 27: 591-604. DOI: 10.1080/03610919808813497

Keselman, H.J., J. Algina, R.K. Kowalchuk and R.D. Wolfinger, 1999a. The analysis of repeated measurements: A comparison of mixed-model Satterthwaite f tests and a nonpooled adjusted degrees of freedom multivariate test. Commun. Stat. Theory Meth., 28: 2967-2999. DOI: 10.1080/03610929908832460

Keselman, H.J., J. Algina, R.K. Kowalchuk, and R.D. Wolfinger, 1999b. A comparison of recent approaches to the analysis of repeated measurements. British J. Math. Stat. Psychol., 52: 63-78. DOI: 10.1348/000711099158964

Khattree, R. and N.D. Naik, 2000. Multivariate Data Reduction and Discrimination with SAS Software. 1st Edn., SAS Institute, Cary, NC SAS Inst., ISBN-10: 1580256961, pp: 574.

Littell, R.C., G.A. Milliken, W.W. Stroup and R.D. Wolfinger, 1999. SAS System for Mixed Models. 1st Edn., Reprint, SAS Institute, Incorporated, Cary, ISBN-10: 1555447791, pp: 633.

Littell, R.C., J. Pendergast and R. Natarajan, 2000. Modelling covariance structure in the analysis of repeated measures data. Stati. Med., 19: 1793-1819. DOI: 10.1002/1097-0258(20000715)19:13<1793::AID-SIM482>3.0.CO;2-Q

McCulloch, C.E., 2003. Generalized Linear Mixed Models. 1st Edn., Illustrated, IMS, Beachwood, ISBN-10: 0940600544, pp: 84.

Núñez-Antón, V. and D.L. Zimmerman, 2000. Modeling nonstationary longitudinal data. Biometrics, 56: 699-705. DOI: 10.1111/j.0006-341X.2000.00699.x

Schwarz, G., 1978. Estimating the dimension of a model. Annal. Stat., 6: 461-464.

Singer, J.D., 1998. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. J. Edu. Behavioral Stat., 24: 323-325. DOI: 10.3102/10769986023004323

Yanosky, II, D.J., 2007. comparability of covariance structures and accuracy of information criteria in mixed model methods for longitudinal data analysis, Published dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy, University of Georgia, Athens, Georgia, USA.