Appendix E: Selecting an Appropriate Covariance Structure

Purpose:
- Motivate the need to model covariance structures in a mixed model.
- Provide a listing and description of commonly used covariance structures.
- Provide guidelines for choosing an appropriate covariance structure.

E.1 Importance of Modeling Covariance (i.e., Temporal or Spatial Dependence in Data):

Many data sets in biological research have spatially and/or temporally correlated observations. An example of spatially correlated data would be most field trails, because adjacent plots will probably respond more alike than will plots that are further apart. An example of temporally correlated data would be any experiment where data for the same variable are recorded on the same experimental unit (plot, container, ...) over time, which is commonly referred to as a 'repeated measures experiment'. If you combine these two examples into a single experiment, field study (Completely Randomized, Randomized Complete Block, Latin Square,....) with repeated measures over time of the same variable on each plot, then you have both spatially and temporally correlated data. This is a common scenario in a variety of studies, including fields, greenhouses, animals, humans, and microarrays.

When data from such experiments lead to **treatment and/or time comparisons**, linear model analyses with correlated errors should be considered. The MIXED model procedure allows for the estimation and testing of differences among treatment and/or time means when the experimental errors are correlated. The purpose of this section is to introduce you to the analysis of temporally and spatially correlated data using the MIXED procedure.

Why is it important to consider the temporal and/or spatial correlation among data? The reason is the effect on sensitivity of tests of differences between means. It is easy to understand this effect by examining the standard error of the difference between treatment means for independent data versus correlated data.

$$S_{\bar{y}_1-\bar{y}_2} = \sqrt{S_e^2 \times (1/n_1 + 1/n_2)}$$

How does the standard error of the difference differ for correlated data?

The standard error of the difference between treatment means for correlated data includes the covariance between measurements that are close enough together in time or space to be correlated.

$$S_{\bar{y}_1-\bar{y}_2} = \sqrt{(S_e^2 - Cov)(1/n_1 + 1/n_2)}$$

Covariance is the unstandardized correlation, that is, it's units of measurement are those for the dependent variable and can vary from minus to plus infinity. When the correlation is zero, the covariance is zero, which corresponds to data that are far enough apart in time or space to be independent. However, as the correlation increases the covariance increases, thus reducing the

variance of the differences for correlated data. As the correlation approaches one, the covariance approaches the $\mathbf{S_e}^2$ and thus the SED for highly correlated data approaches zero. The smaller the SED the greater the precision of the test. There is also a down side. As the correlation increases there is less unique information for estimation of the random variance. This is equivalent to having fewer degrees of freedom for estimation of the random variance. The extreme case would be where the correlation equals one, then additional data provides no new information about variability and no additional degrees of freedom accumulate. Thus if the correlations are close to one the variance of the differences may be small, but so are degrees of freedom and even though precision is high, sensitivity may not be improved due to the small numbers of degrees of freedom.

The Analysis of Repeated Measures "Correlated" Data:

There are several options for analyzing repeated data. Three of these options are listed below, but only the last one will be presented in this class.

-**Run an analysis at each time period.** Since each time period is analyzed separately no statistical hypotheses about time effects can be conducted. In most cases this would be a rather incomplete analysis of the data. However for an accumulative variable this may be the most appropriate analysis. When the dependent variable is cumulative (e.g., height of plants or dry matter accumulation to date), the data at any time contains the information from previous time periods. Even on accumulative variables a repeated measures analysis may be useful if the experimental objectives include about how the accumulation took place over time.

-**Compute a response index** that captures the time-related information for each experimental unit, and conduct the analyses on this index as the dependent variable (eg., Chapter 1, Example 1.3). Common examples of this approach would be the analysis of growth rate, area under the response curve and time to peak response. Such "response indexes" may be parameters resulting from the fitting of non-linear growth curves. Once computed, these parameter values can be used as dependent variables in a linear mixed models analysis.

-**Use repeated measures analysis techniques that estimate the covariance** among residuals after fitting the fixed effects, and use these variance and covariance estimates to compute appropriate standard errors and F-tests for fixed effect hypotheses.

If you wish to use the last approach for the analysis of repeated measures then the MIXED procedure should be used, because a number of different covariance structures are available, one of which is likely to fit your data. Model fitting statistics in the MIXED procedure are useful for determining which covariance structure best describes the random variances and covariances among your repeated measures. It will be necessary to try various covariance structures in order to examine the goodness-of-fit measures for different structures. You will need to become familiar with the available structures (Appendix E.2) in order to select (Appendix E.3) those that might be reasonable for your data.

Covariance measures the degree of association among variables or in this case among repeated measures of the same variable. More specifically repeated measures covariance structures estimate the association among residuals of repeated measurements from the same experimental unit. SAS

provides a number of standard variance and covariance relationships; a few selected ones are illustrated on the following pages. A table summarizing the characteristics of these covariance structures is also provided. See the SAS 9 Help pull-down menu (link provided in C.7) and look under Syntax > Repeated for a comprehensive listing of cavariance structures offered in PROC MIXED.

E.2 A Selected Listing of Covariance Structures:

Important: In all below descriptions, the term "repeated measurement" refers in a broad sense to measurements that may be correlated. Although "repeated measurements" can refer to spatially-related measurements within a local area, these concepts may be most readily comprehended by thinking of "repeated measurement" as a measurement taken sequentially in "time" on the same experimental unit.

VC ***Variance Components*** *(i.e.,* ***Independent*** *or* ***Simple****)*. Equal variances (σ^2_ϵ) on the main-diagonal and zero (0) covariances on the off-diagonals. That is, variances are constant and the residuals are independent across time. This is What is assumed for the standard fixed model ANOVA, but is seldom true with repeated measures. It is considered *simple* because only a single parameter estimate, the pooled variance (σ^2_ϵ) , is required. It is considered *independent* because repeated measurements are not correlated with one another, no matter how close together they are in either time or space.

Repeated Measure	Repeated Measure			
	Time1	Time2	Time3	Time4
Time1	σ^2_ϵ	0	0	0
Time2	0	σ^2_ϵ	0	0
Time3	0	0	σ^2_ϵ	0
Time4	0	0	0	σ^2_ϵ

UN ***Unstructured***. Separate variances for each repeated measurement (σ^2_1, σ^2_2, ...) on the main-diagonal and separate covariances for each pair of repeated measurements (σ_{21}, σ_{31}, ... , σ_{43}) on the off-diagonals. This is essentially the multivariate repeated measures analysis. It is the most complex structure, because a variance is estimated for each repeated measurement and a covariance for each pair of repeated measurements. In the following example, there are t=4 variances (number of repeated measures = t) and six covariances [t(t-1)/2] for a total of ten parameters to be estimated (Note: $\sigma_{ji} = \sigma_{ij}$). **Besides being generally more difficult to solve, the estimates will be less precise as compared to a solution with fewer estimates. Think of this in terms of degrees of freedom. Given a fixed number of degrees of freedom for the random effects, a single variance captures all of those degrees of freedom, while the same degrees of freedom are divided among the estimates if more than one parameter is estimated.**

Repeated Measure	Repeated Measure			
	Time1	Time2	Time3	Time4
Time1	σ^2_1	σ_{12}	σ_{13}	σ_{14}
Time2	σ_{21}	σ^2_2	σ_{23}	σ_{24}
Time3	σ_{31}	σ_{32}	σ^2_3	σ_{34}
Time4	σ_{41}	σ_{42}	σ_{43}	σ^2_4

CS ***Compound Symmetry***. Equal variances ($\sigma^2_\epsilon + \sigma_1$) on the main-diagonal and equal covariances (σ_1) on all off-diagonals (equal correlation). This structure is the simplest repeated measures (i.e., correlated errors) structure. This is the general structure used to analyze data collected according to a split-plot design. This covariance structure require two parameter estimates: σ^2_ϵ and σ_1 . A specific example of a CS covariance structure is a RCBD where $\sigma_1 = \sigma^2_\gamma$.

Repeated Measure	Repeated Measure			
	Time1	Time2	Time3	Time4
Time1	$\sigma^2_\epsilon + \sigma_1$	σ_1	σ_1	σ_1
Time2	σ_1	$\sigma^2_\epsilon + \sigma_1$	σ_1	σ_1
Time3	σ_1	σ_1	$\sigma^2_\epsilon + \sigma_1$	σ_1
Time4	σ_1	σ_1	σ_1	$\sigma^2_\epsilon + \sigma_1$

CSH ***Heterogeneous Compound Symmetry***. This covariance structure allows for unequal variances (σ^2_1, σ^2_2, ...) on the main-diagonal and unequal covariances for all off-diagonals. The magnitude of the covariances is based on the product of the standard deviations ($\sigma_1\sigma_2$, $\sigma_1\sigma_3$, ... , $\sigma_2\sigma_3$, ...) multiplied by a single correlation coefficient (ρ) for the off-diagonal covariances. Thus the correlation over repeated measurements is constant, but the covariances are different depending on the differences in the standard deviations. This covariance structure requires t + 1 parameter estimates. For our example, t=4 and the parameters to be estimated are: σ_1, σ_2, σ_3, σ_4, and ρ.

Repeated Measure	Repeated Measure			
	Time1	Time2	Time3	Time4
Time1	σ^2_1	$\rho\sigma_1\sigma_2$	$\rho\sigma_1\sigma_3$	$\rho\sigma_1\sigma_4$
Time2	$\rho\sigma_2\sigma_1$	σ^2_2	$\rho\sigma_2\sigma_3$	$\rho\sigma_2\sigma_4$
Time3	$\rho\sigma_3\sigma_1$	$\rho\sigma_3\sigma_2$	σ^2_3	$\rho\sigma_3\sigma_4$
Time4	$\rho\sigma_4\sigma_1$	$\rho\sigma_4\sigma_2$	$\rho\sigma_4\sigma_3$	σ^2_4

AR(1) ***First-Order Auto-Regressive***. Equal variances (σ^2_ϵ) on the main-diagonal. While on the off-diagonal "bands", the covariance is the variance times the repeated measures correlation coefficient (ρ) raised to increasing powers (ρ, ρ^2, ρ^3, ...) as the measures become farther separated (in time or space). Since the correlation coefficient is a decimal value, the increasing powers result in decreasing covariances. **Repeated measurements MUST be correctly ordered (in time or space) and an assumption of "equal spacing" between each repeated measurement must be reasonably applicable to your data.** This structure requires the estimation of two parameters (σ_ϵ and ρ) plus it may be necessary to include the experimental unit variance (i.e., Block or Subject effect) on the RANDOM statement since it is not included in the covariance structure.

Repeated Measure	Repeated Measure			
	Time1	Time2	Time3	Time4
Time1	σ^2_ϵ	$\rho\ \sigma^2_\epsilon$	$\rho^2\ \sigma^2_\epsilon$	$\rho^3\ \sigma^2_\epsilon$
Time2	$\rho\ \sigma^2_\epsilon$	σ^2_ϵ	$\rho\ \sigma^2_\epsilon$	$\rho^2\ \sigma^2_\epsilon$
Time3	$\rho^2\ \sigma^2_\epsilon$	$\rho\ \sigma^2_\epsilon$	σ^2_ϵ	$\rho\ \sigma^2_\epsilon$
Time4	$\rho^3\ \sigma^2_\epsilon$	$\rho^2\ \sigma^2_\epsilon$	$\rho\ \sigma^2_\epsilon$	σ^2_ϵ

ARH(1) ***Heterogeneous First-Order Auto-Regressive***. This structure allows for unequal variances (σ^2_1, σ^2_2, ...) on the main-diagonal and unequal covariances on the off-diagonal "bands". The covariances are based on the product of the standard deviations ($\sigma_1\sigma_2$, $\sigma_1\sigma_3$, ... , $\sigma_2\sigma_3$, ...) times the repeated measures coefficient raised to increasing powers (ρ, ρ^2, ρ^3, ...) as the measures become farther separated (in time or space). **Repeated measurements MUST be correctly ordered (in time or space) and an assumption of "equal spacing" between each repeated measurement must be reasonably applicable to your data.** This covariance structure requires the estimation of t + 1 parameters (for our example: σ_1, σ_2, σ_3, σ_4 and ρ) plus it may be necessary to include the experimental unit (i.e., Block or Subject) variance on the RANDOM statement.

Repeated Measure	Repeated Measure			
	Time1	Time2	Time3	Time4
Time1	σ^2_1	$\rho\sigma_1\sigma_2$	$\rho^2\sigma_1\sigma_3$	$\rho^3\sigma_1\sigma_4$
Time2	$\rho\sigma_2\sigma_1$	σ^2_2	$\rho\sigma_2\sigma_3$	$\rho^2\sigma_2\sigma_4$
Time3	$\rho^2\sigma_3\sigma_1$	$\rho\sigma_3\sigma_2$	σ^2_3	$\rho\sigma_3\sigma_4$
Time4	$\rho^3\sigma_4\sigma_1$	$\rho^2\sigma_4\sigma_2$	$\rho\sigma_4\sigma_3$	σ^2_4

TOEP ***Toeplitz***. Equal variances (σ^2_ϵ) on the main-diagonal, equal correlation and covariances (σ_1, σ_2, ...) within each off-diagonal "band", and different correlation and covariances among bands. Because the subscript (i) of σ_i refers to the "distance" between the repeated measurement "periods" (in time or space), **repeated measurements are assumed to be equally spaced and correctly ordered.** TOEP requires t parameter estimates; for our example: σ^2, σ_1, σ_2, and σ_3.

Repeated Measure	Repeated Measure			
	Time1	Time2	Time3	Time4
Time1	σ^2_ϵ	σ_1	σ_2	σ_3
Time2	σ_1	σ^2_ϵ	σ_1	σ_2
Time3	σ_2	σ_1	σ^2_ϵ	σ_1
Time4	σ_3	σ_2	σ_1	σ^2_ϵ

TOEPH ***Heterogeneous Toeplitz***. Although this covariance structure may be difficult to intuitively understand, it is essentially a generalization of the TOEP strucutre. It allows for unequal variances (σ^2_1, σ^2_2, ...) on the main-diagonal, equal correlations within each off-diagonal band, but different covariances within and among off-diagonal bands. The covariances are different because their value is based on the product of the standard deviations ($\sigma_1\sigma_2$, $\sigma_1\sigma_3$, ... , $\sigma_2\sigma_3$, ...) multiplied by the correlation coefficient (ρ_1, ρ_2, ...) for each off-diagonal. **This covariance structure requires correct ordering and assumes equal spacing of the repeated measurements (in time or space).** TOEPH requires t + (t-1) parameter estimates; for this example: σ_1, σ_2, σ_3, σ_4, ρ_1, ρ_2, and ρ_3.

Repeated Measure	Repeated Measure			
	Time1	Time2	Time3	Time4
Time1	σ^2_1	$\rho_1\sigma_1\sigma_2$	$\rho_2\sigma_1\sigma_3$	$\rho_3\sigma_1\sigma_4$
Time2	$\rho_1\sigma_2\sigma_1$	σ^2_2	$\rho_1\sigma_2\sigma_3$	$\rho_2\sigma_2\sigma_4$
Time3	$\rho_2\sigma_3\sigma_1$	$\rho_1\sigma_3\sigma_2$	σ^2_3	$\rho_1\sigma_3\sigma_4$
Time4	$\rho_3\sigma_4\sigma_1$	$\rho_2\sigma_4\sigma_2$	$\rho_1\sigma_4\sigma_3$	σ^2_4

ANTE(1) ***First-Order Ante-Dependence***. Although this covariance structure may be difficult to understand intuitively, it is a generalization of ARH(1) and TOEPH structures whose importance lies in the fact that the assumption of **equal spacing between repeated measurements is NOT required**. This structure allows for unequal variances (σ^2_1, σ^2_2, ...) on the main-diagonal and unequal correlations (ρ_1, ρ_2, ...) and covariances, where ρ_1 is the correlation between repeated measurements 1 and 2, ρ_2 is the correlation between repeated measurements 2 and 3, etc. The magnitude of the covariance depends on the magnitude of

both the correlations and the standard deviations. Repeated measurements must be specified in correct order (in space or time). This variance-covariance structure requires the estimation of t + (t-1) parameters.

Repeated Measure	Repeated Measure			
	Time1	Time2	Time3	Time4
Time1	σ^2_1	$\rho_1\sigma_1\sigma_2$	$\rho_1\rho_2\sigma_1\sigma_3$	$\rho_1\rho_2\rho_3\sigma_1\sigma_4$
Time2	$\rho_1\sigma_2\sigma_1$	σ^2_2	$\rho_2\sigma_2\sigma_3$	$\rho_2\ \rho_3\sigma_2\sigma_4$
Time3	$\rho_2\ \rho_1\sigma_3\sigma_1$	$\rho_2\sigma_3\sigma_2$	σ^2_3	$\rho_3\sigma_3\sigma_4$
Time4	$\rho_3\rho_2\rho_1\sigma_4\sigma_1$	$\rho_3\rho_2\sigma_4\sigma_2$	$\rho_3\sigma_4\sigma_3$	σ^2_4

SP(POW)(*x y*) ***Spatial Power***. This covariance structure is one of several structures available in SAS in which covariances are mathematical functions of Euclidean distances between observed measurements. **These *Spatial* covariance structures can be applied to temporally as well as spatially-related measurements and equal spacing between measurements is NOT required.** The "(*x y*)" is a listing of two numeric variables in your data set that indicate the (x,y)-coordinate location of each observed data value. PROC MIXED uses the actual (x,y) coordinates of your data points to compute the Euclidean distance between each measurement. For our example, assume that the 4 repeated measures are made on the same Experimental Unit at Times: 0 days, 1 day, 3 days, and 7 days. Hence, the d_{ij} are Euclidean distances between times: $d_{12} = d_{21} = 1\text{-}0 = 1$ day, $d_{13} = d_{31} = 3\text{-}0 = 3$ days, $d_{14} = d_{41} = 7\text{-}0 = 7$ days, $d_{23} = d_{32} = 3\text{-}1 = 2$ days, $d_{24} = d_{42} = 7\text{ -}1 = 6$ days, and $d_{34} = d_{43} = 7\text{-}3 = 4$ days. The SP(POW)(*x y*) covariance structure requires that only 2 parameters σ^2 and ρ be estimated.

Details can be found under ... Syntax > Repeated (from your SAS9 menu bar: *Help > SAS Help & Documentation > ... see link in C.7*) regarding the other *Spatial* covariance structures available in PROC MIXED. These structures allow exponential, linear, gaussian, and spherical distributions and anisotrophic (i.e., directionally-changing) covariances.

Repeated Measure	Repeated Measure			
	Time1	Time2	Time3	Time4
Time1	σ^2	$\rho^{d12}\ \sigma^2$	$\rho^{d13}\ \sigma^2$	$\rho^{d14}\ \sigma^2$
Time2	$\rho^{d21}\ \sigma^2$	σ^2	$\rho^{d23}\ \sigma^2$	$\rho^{d24}\ \sigma^2$
Time3	$\rho^{d31}\ \sigma^2$	$\rho^{d32}\ \sigma^2$	σ^2	$\rho^{d34}\ \sigma^2$
Time4	$\rho^{d41}\ \sigma^2$	$\rho^{d42}\ \sigma^2$	$\rho^{d43}\ \sigma^2$	σ^2

E.3 A Process for Choosing a Suitable Covariance Structure[1]:

To find a suitable covariance structure, try the following:

1) Try running the TYPE = UN (i.e. "unstructured" covariance) first. This is the most complex structure and often fails to run with few replications or when there are many repeated measures. This is generally not the one to use, but examination of the patterns in this covariance matrix may often suggest a simpler structure that fits well.

2) Next, run the TYPE = CS (i.e., compound symmetry) structure. Compound symmetry is the simplest repeated measures structure. In some cases it may be clear that a heterogeneous variance structure will be needed. In such cases, you my choose TYPE = CSH as the simplest reasonable structure.

3) Now run other covariance structures (Appendix E.2) that are reasonable based on: any patterns seen in the unstructured matrix, the biology, and on your knowledge of the experiment (spacing of time periods relative to the biology of the organism).

4) If none of the above covariance structures produce a significant (i.e., Pr > ChiSq < .05) test for the Null Model Likelihood Ratio Test, this indicates there is no correlation present in the model's error structure. Use TYPE=VC to fit an independent covariance structure.

The Null Model Likelihood Ratio Test is a test to determine whether the covariance structure specified in the TYPE= statement fits the data significantly better than a model that assumes no correlation (i.e., independence ≡ TYPE=VC). Typically, more than one of the candidate covariance structures will yield a significant Null Model Likelihood Ratio Test. The most appropriate covariance structure for the data is the structure with the smallest Akaike's Information Criteria (AICC) value. When two or more covariance structures yield a similarly small AICC value, let parsimony guide you to choose the model with the fewest number of parameters (i.e., distinct covariance values). Creating a diagnostics graph (Appendix B.2.5 and Appendix F.6.c.v.) will also assist in making the best selection.

[1]See Chapter 5 for an illustrative example of this covariance selection process.

Note: Significance of the F-test **CANNOT** be used as a criterion for selecting the covariance structure for a model.