The analysis of very small samples of repeated measurements I: An adjusted sandwich estimator

Simon S. Skene[a*†] **and Michael G. Kenward**[b]

The statistical analysis of repeated measures or longitudinal data always requires the accommodation of the covariance structure of the repeated measurements at some stage in the analysis. The general linear mixed model is often used for such analyses, and allows for the specification of both a mean model and a covariance structure. Often the covariance structure itself is not of direct interest, but only a means to producing valid inferences about the response. Existing methods of analysis are often inadequate where the sample size is small. More precisely, statistical measures of goodness of fit are not necessarily the right measure of the appropriateness of a covariance structure and inferences based on conventional Wald type procedures do not approximate sufficiently well their nominal properties when data are unbalanced or incomplete. This is shown to be the case when adopting the Kenward-Roger adjustment where the sample size is very small. A generalization of an approach to Wald tests using a bias adjusted empirical sandwich estimator for the covariance matrix of the fixed effects from generalized estimating equations is developed for Gaussian repeated measurements. This is shown to attain the correct test size but has very low power.

Keywords: covariance matrix; cross-over trials; empirical sandwich estimator; Kenward-Roger adjustment; repeated measures; small samples

1. Introduction

The purpose of this paper and the following companion paper [1] is to examine methods of analysis which are appropriate for very small samples of repeated measurements. These issues are highlighted in Brammer [2], who suggests using the general linear mixed model as an appropriate modelling framework for ascending dose design studies, with reference to two experiments involving isolated tissues and organs from guinea pigs and rats. In one of these experiments, an attempt is made to estimate a covariance matrix for the seven repeated measurements on an organ from data relating to just three subjects, and in the other a covariance matrix from four repeated measurements from each of 12 subjects. Although

[a] *Department of Economics, Mathematics and Statistics, Birkbeck, London, UK.*
[b] *Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK.*
[*] *Correspondence to: Simon S. Skene, Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK.*
[†] *E-mail: s.skene@bbk.ac.uk*

the first of these examples may be thought of as an extreme case, this type of problem is typical of many encountered in early stage drug development, where studies are routinely undertaken using very small samples due to costs and the desire to minimize unnecessary animal experimentation. A particular example is in the area of safety pharmacology, where investigations are undertaken into the effects of test compounds on the vital functions of the central nervous, cardiovascular and respiratory systems prior to trials involving human subjects. Such experiments are carried out under 'good laboratory practice' to give an indication of potentially adverse effects, but can also indicate possible clinical benefits. It is not unusual to use four to eight subjects in longitudinal and cross-over experiments.

Repeated measurements are typically mutually correlated, and commonly have variances that change over time. Such covariance structures need to be accommodated in the analysis and there are several ways in which this may be done. The problem with very small samples is that this structure will be poorly estimated and any aspects of the analysis that depend on its estimation will be affected by this. Reducing the number of covariance parameters, that is choosing a more parsimonious structure, would appear to provide one way of reducing this problem, but to avoid invalidity of subsequent inference the chosen structure must be of adequate fit. Unfortunately, assessment of fit of a covariance structure is also undermined by the lack of information in such small samples. In this paper we show how these concerns apply directly to conventional techniques for analyzing continuous repeated measurements that are based on a full likelihood analysis with the multivariate linear model (see [3] for example). In the light of these problems we consider, as an alternative, the use of ordinary least squares that ignores the covariance structure in the estimation step, combined with a sandwich estimator of error that corrects for the covariance structure. Our sandwich estimator combines two different small sample adjustments for such estimators suggested by Pan and Wall [4] and Mancl and DeRouen [5] respectively. It is shown that this estimator leads to a test statistic with good nominal size, but unfortunately with extremely low power.

The paper is arranged as follows. In Section 2 we review existing methods for the analysis of repeated measurements through the multivariate linear model, including the choice of covariance model and small-sample inference. In Section 3 we develop the new adjustment for the empirical sandwich estimator of the fixed effects for Gaussian data. The performance of procedures based on these methods is assessed in Section 4 through a series of simulation studies, and discussed in Section 5.

2. A review of existing methods

The multivariate Gaussian linear model is often adopted for the analysis of repeated measurements [3]. Let $\mathbf{y}_i$ $(T_i \times 1)$ be the response vector from the ith of n subjects, then the model has the following general form

$$\mathbf{y}_i \sim \mathrm{N}\left(\mathbf{X}_i\boldsymbol{\beta}; \boldsymbol{\Sigma}_i\right), \quad i = 1, \ldots, n, \tag{1}$$

for $\boldsymbol{\beta}$, $(p \times 1)$, the vector of fixed effect parameters, $\mathbf{X}_i$, $(T_i \times p)$, the design/covariate matrix, and $\boldsymbol{\Sigma}_i$ the $(T_i \times T_i)$ covariance matrix for this subject. Depending on the setting, the covariance matrix can in principle take many forms, including those induced by a random effects structure. Defining $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$, setting $\boldsymbol{\Sigma} = \text{block-diagonal}\,\{\boldsymbol{\Sigma}_i\}$ we have the equivalent expression for the whole data set:

$$\boldsymbol{Y} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}; \boldsymbol{\Sigma}\right). \tag{2}$$

Usually restricted (or residual) maximum likelihood (REML) [6] is used to estimate the parameters of this model, and is typically preferred to maximum likelihood in small samples because it takes account of the loss of degrees of freedom from estimating the fixed effects. $\hat{\boldsymbol{\Sigma}}$ is chosen to maximise the REML log-likelihood:

$$l_R(\boldsymbol{\Sigma}) \propto -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log|\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{y}^T\mathbf{H}_{\boldsymbol{\Sigma}}\mathbf{y} \tag{3}$$

where $\mathbf{H}_{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}$, and the REML estimator of β is the generalized least squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}. \tag{4}$$

This formulation is dependent on a choice of structure for $\boldsymbol{\Sigma}_i$, but there is no universally satisfactory approach to finding this. When sample sizes are not too small, and there is a limited fixed set of measurement times, the unstructured matrix is usually a satisfactory choice (see for example [7] Section 5.6). When times are not common among subjects or samples sizes are small then an alternative approach is to impose a particular pattern on the variances and correlations, either directly, or through a formulation with random effects. Although some have suggested the use of stationary covariance structures, such as compound symmetry and first-order autoregressive, the non-stationarity of repeated measurements, which is common outside the cross-over setting, precludes such patterns and has led to the introduction of non-stationary models such as antedependence (see, for example [8]). Wolfinger [9] reviews a large number of such proposals and suggests the following strategy based on Diggle [10] for model selection.

(1) Use graphical analysis to select an initial mean model.

(2) Select initial covariance structures by any relevant means.

(3) Use formal statistical techniques to compare and select an appropriate structure. Assuming the chosen covariance structure, reduce the mean model if necessary.

A number of criteria can be used to compare the fit of such covariance structures. For instance, (reduced) likelihood ratio tests can be used for nested models to test whether the additional parameters of the more complicated covariance structure give a significant improvement to the fit of the model. That is,

$$2\{l_R(\boldsymbol{\Sigma}_1) - l_R(\boldsymbol{\Sigma}_2)\} \sim \chi^2_{q_1 - q_2} \tag{5}$$

where $\mathbf{\Sigma}_2$ is nested within $\mathbf{\Sigma}_1$ and q_i are the corresponding number of covariance parameters for each model. Also, Akaike's information criterion (AIC, [11]) and Schwarz's Bayesian criterion (BIC, [12]) are both log-likelihood measures penalized by the number of parameters, which can be used to make a direct comparison between models which fit the same fixed effects. These can be expressed in terms of minus twice the maximized log REML (*i.e.* smaller is better) as

$$\begin{aligned} \text{AIC}_R &: -2l_R + 2q \\ \text{BIC}_R &: -2l_R + q\log(n) \end{aligned} \tag{6}$$

where q is the number of covariance parameters in the selected model, and n is the number of effective subjects.

An alternative route to the development of covariance structures is to model directly the behaviour of subjects through a random coefficients model, and to use the structure induced by this. (See, for example, [13]). While attractive in principle the success of such an approach is critically dependent on the appropriate choice of subject level model, and this can be difficult in practice.

Graphical and descriptive methods such as the semi-variogram [10] can point to likely classes of covariance structure, but comparison of the fit of a number of models is usually necessary for any given problem. This can be time consuming, and commonly the choice of structure is not consistent among similar types of data set and, most importantly in the current setting, will not work well with very small samples.

Once an appropriate covariance structure has been estimated, the resulting fixed effects parameters obtained from (4) are asymptotically distributed as $\hat{\boldsymbol{\beta}} \sim \mathbf{N}(\boldsymbol{\beta}, \mathbf{\Phi})$, where

$$\mathbf{\Phi} = (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}. \tag{7}$$

Inferences about the fixed effects made via the general linear hypothesis, $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{L}$ is an $(l \times p)$ fixed matrix whose rows are linearly independent, may be tested using the approximate Wald statistic with $\hat{\mathbf{\Phi}} = (\mathbf{X}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^{-1}$ replacing the true value:

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{L}^T(\mathbf{L}\hat{\mathbf{\Phi}}\mathbf{L}^T)^{-1}\mathbf{L}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_l \tag{8}$$

To accommodate the uncertainty in $\hat{\mathbf{\Phi}}$ a Wald F-statistic is usually preferred:

$$F = l^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{L}^T(\mathbf{L}\hat{\mathbf{\Phi}}\mathbf{L}^T)^{-1}\mathbf{L}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{9}$$

where F is assumed to have a null $\text{F}_{l,v}$ distribution and the denominator degrees of freedom v are estimated from the data. One simple choice is the residual degrees of freedom $N - \text{rank}(\mathbf{X})$, where N is the total number of observations, although there are a number of alternatives which all attempt to accommodate for the finite sample behaviour of the Wald statistic. In a very general approach to this, Kenward and Roger [14, 15] first replace $\hat{\mathbf{\Phi}}$ in (9) by an adjusted covariance matrix for the fixed effects parameters that accommodates the uncertainty in $\hat{\mathbf{\Sigma}}$, and then approximate a scaled F-statistic, $F^* = \lambda F$, by an $\text{F}_{l,v}$

distribution, with λ and v estimated from the data through matching the first two moments of F^* and the approximating F-distribution. The approximation has the advantage that it recovers the correct values of λ and v in those special cases in which the Wald statistic has an exact F-distribution, namely of Hotelling's T^2 and balanced random effects analysis of variance. This approximation also recovers Satterthwaite's adjusted degrees of freedom [16] for the one degree of freedom Wald test (*i.e.* $l = 1$) for which λ is identically one. A number of alternatives for calculating the denominator degrees of freedom for the Wald F-statistic are reviewed by [17], who recommend the use of the Kenward-Roger method. They show it outperforms the other options in a simulation study, but note that when the covariance structure is complicated and the sample size is small this method can lead to inflated type 1 error rates.

In the likelihood based approach the estimated covariance matrix $\hat{\mathbf{\Sigma}}$ plays two roles: it influences directly the estimator of $\boldsymbol{\beta}$ and it is used to calculate the precision of this, and its great imprecision in very small samples can undermine either or both of these roles. For this reason we may want to reduce or remove its impact on either of these. One route to this is to ignore the estimated covariance structure entirely for the purposes of estimation and subsequently obtain an appropriate estimate of precision. In doing this we have removed its direct impact on estimation, but not from the estimate of precision. Such an approach is well known and amounts to a linear model implementation of generalised estimating equations (GEE's) with the robust or empirical variance estimator, [18], also known as the 'sandwich' estimator. The estimator can be written

$$\begin{aligned} \hat{\boldsymbol{\beta}}_W &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} \\ &= (\sum_{i=1}^{n}\mathbf{X}^T\mathbf{W}_i\mathbf{X}_i)^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{W}_i\mathbf{y}_i \end{aligned} \tag{10}$$

for $\mathbf{W}^{-1} = \text{block-diagonal}\{\mathbf{W}_i^{-1}\}$ a fixed working covariance matrix. This is unbiased for $\boldsymbol{\beta}$. The corresponding empirical sandwich estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}_W$ can be written

$$\mathbf{V}_S = \hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}_W) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{W}_i\hat{\mathbf{\Sigma}}_i\mathbf{W}_i\mathbf{X}_i\right)(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \tag{11}$$

for $\mathbf{\Sigma}_i = \mathrm{V}(\mathbf{y}_i)$. $\mathbf{V}_S$ will be a consistent estimator of $\mathrm{V}(\hat{\boldsymbol{\beta}}_W)$ provided $\hat{\mathbf{\Sigma}}_i$ is consistent for $\mathrm{V}(\mathbf{y}_i)$ and the usual choice for this is $\hat{\mathbf{\Sigma}}_i = (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^T$ The advantage of this estimator is that a poor choice of $\mathbf{W}$ will not affect the validity of inferences about $\boldsymbol{\beta}$. However it should be noted that this approach causes the fixed effects variances to reflect the observed variances and covariances rather than those of any imposed 'structure'. In the simple case of $\mathbf{W} = \mathbf{I}$, then we have the ordinary least squares estimate of $\boldsymbol{\beta}$ but with the standard errors of the estimates adjusted to account for the observed covariance structure. [19] suggest the use of this approach whenever the mean response is of primary interest in an analysis and limited time is available to determine an appropriate covariance model, or in pharmaceutical trials where statistical methods have to be specified in advance in the protocol.

A disadvantage of the 'robust' method outlined above is that the resulting tests based on asymptotic null chi-squared distributions are known to be unreliable in small samples. That is, the consistency provided by this estimator comes at the price of increased variability. [20] suggest that the robust approach is suitable only when the data come from 'many experimental units', although several authors have attempted to correct for this in certain situations. (See, for example, [21]). One such approach is easily extended to the general linear hypothesis for Gaussian repeated measurements, and is developed in the next section.

3. An adjusted sandwich estimator

Pan and Wall [4] propose an F-test for the Wald statistic which takes account of the variability in the sandwich estimator. They describe their approach as being analogous to that of testing the mean of a normal distribution with unknown variance where a t-test is preferred over a z-test. Their approach is developed in the GEE setting in terms of the simple hypothesis H_0: $\boldsymbol{\beta} = \mathbf{0}$, although they note the possibility for its extension to the general linear hypothesis H_0: $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, which is presented below.

Consider the Wald statistic under the null hypothesis,

$$W = (\mathbf{L}\hat{\boldsymbol{\beta}})^T(\mathbf{L}\mathbf{V}_S\mathbf{L}^T)^{-1}\mathbf{L}\hat{\boldsymbol{\beta}} = v(\mathbf{L}\hat{\boldsymbol{\beta}})^T(v\mathbf{L}\mathbf{V}_S\mathbf{L}^T)^{-1}\mathbf{L}\hat{\boldsymbol{\beta}} \tag{12}$$

where $\mathbf{L}\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{0}, \mathbf{L}\mathbf{V}_S\mathbf{L}^T)$. If we assume further that $v\mathbf{L}\mathbf{V}_S\mathbf{L}^T$ has a scaled Wishart distribution, that is $v\mathbf{L}\mathbf{V}_S\mathbf{L}^T \sim \mathrm{W}_l(v, \mathbf{L}\mathrm{V}(\hat{\boldsymbol{\beta}})\mathbf{L}^T)$, then it follows that W has the same distribution as Hotelling T^2. That is,

$$\frac{v-l+1}{vl}W \sim \mathrm{F}_{l,v-l+1} \tag{13}$$

where v is chosen from the data to match $\widehat{\mathrm{V}}\{\mathrm{vec}(v\mathbf{L}\mathbf{V}_S\mathbf{L}^T)\} = v^2\widehat{\mathrm{V}}\{\mathrm{vec}(\mathbf{L}\mathbf{V}_S\mathbf{L}^T)\}$ from the Wishart distribution with an empirically based estimator.

From the properties of the Wishart distribution, we have

$$\mathrm{V}\{\mathrm{vec}(v\mathbf{L}\mathbf{V}_S\mathbf{L}^T)\} = v(\mathbf{I}_{l^2} + \mathbf{K})\left[\{\mathbf{L}\mathrm{V}(\hat{\boldsymbol{\beta}})\mathbf{L}^T\} \otimes \{\mathbf{L}\mathrm{V}(\hat{\boldsymbol{\beta}})\mathbf{L}^T\}\right] \tag{14}$$

where $\mathbf{K}$ is the 'commutative' matrix and $\otimes$ is the Kronecker product operator. Also, since $\mathbf{V}_S$ is a consistent estimator of the covariance matrix of the fixed effects parameters, $v\hat{\boldsymbol{\Omega}}$ is a consistent estimator of $\mathrm{V}\{\mathrm{vec}(v\mathbf{L}\mathbf{V}_S\mathbf{L}^T)\}$ where

$$\hat{\boldsymbol{\Omega}} = (\mathbf{I}_{l^2} + \mathbf{K})\{(\mathbf{L}\mathbf{V}_S\mathbf{L}^T) \otimes (\mathbf{L}\mathbf{V}_S\mathbf{L}^T)\} \tag{15}$$

Let $\mathbf{P}_i = \mathrm{vec}(\mathbf{X}_i^T\mathbf{W}_i\hat{\boldsymbol{\Sigma}}_i\mathbf{W}_i\mathbf{X}_i)$, where as before $\hat{\boldsymbol{\Sigma}}_i = (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^T$. Then the

covariance matrix of the mean vector $\mathbf{Q} = \sum_{i=1}^{n} \mathbf{P}_i/n$ is estimated empirically by

$$\begin{aligned} \hat{\mathbf{T}} &= \sum_{i=1}^{n} (\mathbf{P}_i - \bar{\mathbf{P}})(\mathbf{P}_i - \bar{\mathbf{P}})^T/n(n-1) \\ &= \frac{1}{n^2}\widehat{\mathrm{V}}(\mathbf{P}_i) \end{aligned} \tag{16}$$

where $\bar{\mathbf{P}} = \sum_{i=1}^{n} \mathbf{P}_i/n$. This follows since $\sum_{i=1}^{n} \mathbf{P}_i/n$ is itself an unbiased estimator of $\mathbf{Q}$. Then, the covariance matrix of $\mathrm{vec}(\mathbf{L}\mathbf{V}_S\mathbf{L}^T)$,

$$\mathrm{V}\{\mathrm{vec}(\mathbf{L}\mathbf{V}_S\mathbf{L}^T)\} = (\mathbf{L}\mathbf{V}_M \otimes \mathbf{L}\mathbf{V}_M) \sum_{i=1}^{n} \mathbf{P}_i$$

where $\mathbf{V}_M = \hat{\mathbf{\Phi}} = (\mathbf{X}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^{-1}$ is the model-based estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$. This quantity is given (empirically) by

$$\begin{aligned} \hat{\mathbf{\Psi}} = \widehat{\mathrm{V}}\{\mathrm{vec}(\mathbf{L}\mathbf{V}_S\mathbf{L}^T)\} &= (\mathbf{L}\mathbf{V}_M \otimes \mathbf{L}\mathbf{V}_M)\widehat{\mathrm{V}}(\mathbf{P}_i)(\mathbf{L}\mathbf{V}_M \otimes \mathbf{L}\mathbf{V}_M)^T \\ &= n^2(\mathbf{L}\mathbf{V}_M \otimes \mathbf{L}\mathbf{V}_M)\hat{\mathbf{T}}(\mathbf{L}\mathbf{V}_M \otimes \mathbf{L}\mathbf{V}_M)^T \end{aligned} \tag{17}$$

and v is chosen to match as closely as possible the quantities $v\hat{\mathbf{\Omega}}$, the estimated covariance matrix of $v\mathbf{L}\mathbf{V}_S\mathbf{L}^T$ under the Wishart assumption with $v^2\hat{\mathbf{\Psi}}$, its empirically calculated estimate.

The solution favoured by Pan and Wall is to find v to minimise the sum of squared errors between $v\mathrm{vec}(\hat{\mathbf{\Psi}})$ and $\mathrm{vec}(\hat{\mathbf{\Omega}})$. Writing, $\mathbf{a} = \mathrm{vec}(\hat{\mathbf{\Psi}})$ and $\mathbf{b} = \mathrm{vec}(\hat{\mathbf{\Omega}})$, the least squares estimate of v minimises $(v\mathbf{a} - \mathbf{b})^T(v\mathbf{a} - \mathbf{b})$, so that differentiating and setting equal to zero, we find

$$2v\mathbf{a}^T\mathbf{a} - 2\mathbf{a}^T\mathbf{b} = 0$$

which leads to

$$v = \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}} = \frac{\mathrm{tr}(\hat{\mathbf{\Psi}}\hat{\mathbf{\Omega}})}{\mathrm{tr}(\hat{\mathbf{\Psi}}^2)} \tag{18}$$

Pan and Wall compare their adjusted F-statistic for a Wald test using the sandwich estimator with the usual χ^2_l distribution, and show in the context of a logistic model for a binary response that their adjusted test reduces the inflated test sizes achieved using a χ^2 test to nominal levels. They note further that it is possible to combine their approach with a bias correction such as that of Mancl and DeRouen [5]. This is achieved by simply replacing $\mathbf{r_i} = (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$ with $(\mathbf{I} - \mathbf{H}_i)^{-1}\mathbf{r}_i$ in $\mathbf{P}_i$ from (16), where

$$\mathbf{H}_i = \mathbf{X}_i\mathbf{V}_M\mathbf{X}_i^T\mathbf{W}_i \tag{19}$$

and using the bias-adjusted estimator to determine v. This results in a Wald statistic that accounts for both the bias and variability of the sandwich estimator. However, they find in simulations that there is no need for such a combination as their adjusted statistic alone is somewhat conservative for simple hypotheses in the GEE setting, giving type 1 error rates below the nominal level. The simulations of the next section show that in the context of general hypotheses concerning Gaussian observations, the bias correction is necessary in addition to the Pan and Wall adjustment to adequately control the type 1 error rate.

Table I. A cross-over design for five treatments (A)-(E).

	Period				
Subject	1	2	3	4	5
1	A	B	C	D	E
2	B	D	A	E	C
3	D	E	B	C	A
4	E	C	D	A	B
5	C	A	E	B	D
6	E	D	C	B	A
7	C	E	A	D	B
8	A	C	B	E	D
9	B	A	D	C	E
10	D	B	E	A	C

4. Simulation studies

Three study designs are considered to investigate the effects of missing values and model design on choice of covariance model and inference for repeated measurements in a small sample setting.

(A) A simple repeated measures experiment, with 10 subjects randomly allocated to two treatment groups (of equal size), and a response recorded for each subject at each of five time points.

(B) As design (A), but with missing values. One subject in each treatment group drops out at some random time following the first observation.

(C) A five treatment-five period cross-over trial, with 10 subjects allocated randomly to treatments according to Table I, using a pair of Williams' squares.

Under each study design, 1000 data samples are independently generated from a Gaussian distribution with zero mean and each of seven underlying covariance structures; identity, compound symmetry, AR1 (low and high correlation), first order antedependence, and two 'unstructured' forms. These structures are shown in the Appendix. The two unstructured forms are chosen to have patterns far removed from that usually fitted by the structured forms, such as correlations which increase by lag. One of these is found by adoption of a quadratic random effects model with unusual parameter values. The number of covariance parameters to be estimated by each model are then 1 (identity), 2 (compound symmetry, AR1), 9 (antedependence), and 15 (unstructured).

In designs (A) and (B) a saturated means model is fitted, but for the cross-over design (C) the mean model comprises an intercept, treatment and period effects (9 parameters). REML estimates of the mean models and covariance structures are used. In terms of inference, interest in designs (A) and (B) centres on the treatment/time interaction, but in design (C) the appropriate null hypothesis is that of no treatment effect.

Table II compares the use of (reduced) likelihood ratio tests and $\text{AIC}_R/\text{BIC}_R$ measures to identify the correct covariance structure for the simulated data arising from the three study designs. Table III assesses the effectiveness of Wald tests, in terms of type 1 error rate (size) and power, to determine departures from the null hypothesis for the same data.

4.1. Choice of covariance structure

As the unstructured covariance form will always give the best fit to the data, the proportion of significant tests from reduced likelihood ratio tests of the estimated 'true' covariance structure against an unstructured form indicates the actual test size of the test for detecting a more parsimonious structure where an appropriate one exists. This can be thought of as a goodness-of-fit test for a given structure.

$$H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}_S \quad \text{vs} \quad H_1 : \mathbf{\Sigma} = \mathbf{\Sigma}_{UN}$$

Table II shows that the reduced likelihood ratio tests of the 'true' structure against the unstructured do not not have nominal size. That is, where the data are generated from an underlying 'true' structure of few parameters, the likelihood ratio test is not effective at discriminating between this and an unstructured model in these settings. The actual sizes from the simulations are around 25% for design (A) and in excess of 40% for designs (B) and (C) compared to a nominal size of 5%. These sizes are inflated further in the case of the estimated antedependence structure, which shares variances and lag 1 correlations with the unstructured model. Test sizes are inflated to an even greater extent where the nominal size is 1%.

The measures AIC_R and BIC_R have no nominal properties against which to assess their behaviour, so need to be assessed in an absolute sense. For design (A), the measures AIC_R and BIC_R pick out the 'true' structure on the majority of occasions, but with a greater spread towards structures which are close to the 'true' structure. For designs (B) and (C), the spread is increased, indicating the reduced effectiveness of these measures in discriminating between structures.

4.2. Testing of fixed effects

Table III shows greatly inflated type 1 error rates for the χ^2 and F-Wald statistics in design (A), but the Kenward-Roger (KR) adjustment is seen to correct the size at the nominal level of 5% when either the unstructured or 'true' covariance estimate is adopted. If the 'true' structure with few parameters is estimated, then it is more powerful in detecting a significant difference in the interaction. For design (B) we see that when data are missing, the Kenward-Roger adjustment leads to inflated type 1 error rates of around 7.5% when the unstructured or ('true') antedependence structure are used. When the estimate of the 'true'

Table II. Summary of results from 1000 simulations of study designs (A)-(C): Choice of covariance structure.

Underlying 'true' covariance structure	Proportion of significant tests: reduced LRTs 'true' model v's unstr. df	at 5%	at 1%	Proportion chosen by AIC_R (BIC_R) (No. of covariance parameters) UN (15)	ID (1)	CS (2)	AR1 (2)	ANTE (9)
Study design (A)								
Identity	14	0.240	0.099	0.101 (0.047)	0.654 (0.748)	0.117 (0.105)	0.092 (0.085)	0.036 (0.015)
Compound symmetry	13	0.229	0.097	0.110 (0.042)	0.046 (0.060)	0.705 (0.751)	0.116 (0.134)	0.023 (0.013)
AR1 (ρ=0.2)	13	0.254	0.108	0.108 (0.043)	0.435 (0.547)	0.142 (0.131)	0.265 (0.254)	0.050 (0.025)
AR1 (ρ=0.8)	13	0.249	0.122	0.131 (0.064)	0.000 (0.000)	0.071 (0.078)	0.734 (0.819)	0.064 (0.039)
Antedependence	6	0.277	0.142	0.319 (0.220)	0.000 (0.000)	0.000 (0.000)	0.001 (0.002)	0.680 (0.778)
Unstructured	—	—	—	0.985 (0.962)	0.000 (0.003)	0.006 (0.019)	0.003 (0.008)	0.006 (0.008)
Unstr. (QRE)	—	—	—	0.528 (0.415)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.472 (0.585)
Study design (B)								
Identity	14	0.393	0.202	0.197 (0.117)	0.567 (0.671)	0.102 (0.092)	0.095 (0.099)	0.039 (0.021)
Compound symmetry	13	0.403	0.211	0.241 (0.142)	0.054 (0.081)	0.564 (0.629)	0.118 (0.134)	0.023 (0.014)
AR1 (ρ=0.2)	13	0.398	0.204	0.204 (0.110)	0.436 (0.531)	0.125 (0.127)	0.200 (0.207)	0.035 (0.025)
AR1 (ρ=0.8)	13	0.379	0.186	0.207 (0.124)	0.000 (0.001)	0.077 (0.093)	0.655 (0.747)	0.061 (0.035)
Antedependence	6	0.476	0.275	0.505 (0.419)	0.000 (0.004)	0.001 (0.002)	0.001 (0.004)	0.493 (0.571)
Unstructured	—	—	—	0.971 (0.924)	0.003 (0.010)	0.014 (0.042)	0.003 (0.010)	0.009 (0.014)
Unstr. (QRE)	—	—	—	0.628 (0.533)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.372 (0.467)
Study design (C)								
Identity	14	0.410	0.262	0.255 (0.173)	0.531 (0.634)	0.081 (0.074)	0.096 (0.092)	0.037 (0.027)
Compound symmetry	13	0.468	0.287	0.315 (0.228)	0.014 (0.022)	0.542 (0.619)	0.105 (0.120)	0.024 (0.011)
AR1 (ρ=0.2)	13	0.468	0.289	0.300 (0.201)	0.379 (0.477)	0.091 (0.092)	0.183 (0.201)	0.047 (0.029)
AR1 (ρ=0.8)	13	0.485	0.289	0.317 (0.221)	0.000 (0.001)	0.054 (0.061)	0.568 (0.679)	0.061 (0.038)
Antedependence	6	0.534	0.371	0.557 (0.472)	0.000 (0.000)	0.004 (0.005)	0.001 (0.005)	0.438 (0.518)
Unstructured	—	—	—	0.986 (0.968)	0.000 (0.001)	0.009 (0.021)	0.003 (0.005)	0.002 (0.005)
Unstr. (QRE)	—	—	—	0.662 (0.591)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.338 (0.409)

Designs (A) and (B) are simple repeated measures designs of 10 subjects in two equally sized treatment groups measured at 5 time points, (one subject in each group may drop out at some random time following the first observation in Design (B)). Design (C) is a 5 treatment-5 period cross-over experiment of 10 subjects allocated to treatment sequence according to a Williams square. Table gives proportion of significant reduced likelihood ratio tests (RLRTs) for goodness-of-fit for the covariance structure at the fixed levels of $\alpha = 0.05$ and $\alpha = 0.01$. Also, the proportion of structures selected by Akaike's Information criterion (AIC_R) and Schwarz's Bayesian criterion (BIC_R).

Table III. Summary of results from 1000 simulations of study designs (A)-(C): Wald tests using χ^2 and F distributions, and the Kenward-Roger adjustment.

Underlying 'true' covariance structure	Covariance estimate	Proportion of significant test results			
		Size			Power
		χ^2	F	KR	(KR)
Study design (A)					
Identity	Unstructured	0.329	0.290	0.053	0.735
	Identity	0.064	0.048	0.048	0.980
Compound symmetry	Unstructured	0.350	0.306	0.045	0.747
	Comp. sym.	0.069	0.051	0.046	0.985
AR1 (ρ=0.2)	Unstructured	0.352	0.319	0.050	0.735
	AR1	0.095	0.070	0.052	0.978
AR1 (ρ=0.8)	Unstructured	0.344	0.318	0.056	0.775
	AR1	0.074	0.058	0.033	0.984
Antedependence	Unstructured	0.332	0.300	0.058	0.760
	Antedependence	0.208	0.179	0.069	0.945
Unstructured	Unstructured	0.348	0.316	0.045	0.768
Unstr. (QRE)	Unstructured	0.333	0.300	0.048	0.740
Study design (B)					
Identity	Unstructured	0.468	0.432	0.073	0.633
	Identity	0.062	0.044	0.044	0.958
Compound symmetry	Unstructured	0.437	0.411	0.074	0.632
	Comp. sym.	0.080	0.050	0.047	0.950
AR1 (ρ=0.2)	Unstructured	0.461	0.419	0.089	0.637
	AR1	0.094	0.063	0.050	0.976
AR1 (ρ=0.8)	Unstructured	0.463	0.412	0.078	0.637
	AR1	0.075	0.051	0.031	0.966
Antedependence	Unstructured	0.487	0.454	0.084	0.645
	Antedependence	0.261	0.220	0.079	0.858
Unstructured	Unstructured	0.432	0.399	0.077	0.640
Unstr. (QRE)	Unstructured	0.360	0.338	0.046	0.516
Study design (C)					
Identity	Unstructured	0.913	0.900	0.654	—
	Identity	0.065	0.048	0.048	—
Compound symmetry	Unstructured	0.921	0.912	0.690	—
	Comp. sym.	0.077	0.058	0.054	—
AR1 (ρ=0.2)	Unstructured	0.910	0.900	0.660	—
	AR1	0.085	0.061	0.041	—
AR1 (ρ=0.8)	Unstructured	0.914	0.901	0.672	—
	AR1	0.062	0.042	0.024	—
Antedependence	Unstructured	0.943	0.938	0.712	—
	Antedependence	0.410	0.382	0.137	—
Unstructured	Unstructured	0.919	0.910	0.677	—
Unstr. (QRE)	Unstructured	0.926	0.913	0.698	—

Table gives proportion of type 1 errors (Size) and Power using a fixed level of $\alpha = 0.05$. Power is calculated by adding a linear effect in time to one treatment group to achieve approximately 75% using a Kenward-Roger adjustment with the covariance model indicated in Designs (A) and (B) with no missing values. KR - Kenward-Roger adjusted test. Power is not assessed where the test size is not fixed at the nominal level (—).

structure is based on few parameters however, the adjustment still generally corrects the test size to the nominal value.

The loss of nominal properties in these tests when using an unstructured covariance model for data with missing values is marked, although when only two subjects drop out, the loss in data is at most 8 observations out of a possible 50. Allowing a third subject to drop out (from either treatment group) is found to inflate this test size further, to around 15%, although as the amount of missing data increases so do the problems of convergence in finding the REML estimates. With the introduction of missing data, these small sample adjustments no longer lead to exact tests. That is, the Kenward-Roger adjustment for all models is now based on an approximate Taylor series expansion, which may not be appropriate where there are a large number of parameters to be estimated from such small data sets. Also, the covariance structure of the data now directly effects the estimates of the mean parameters as well as their standard errors.

Once the size of the test has been fixed at the nominal level, it is appropriate to consider power. Here this is achieved by adding terms which are linearly increasing in time to the responses of one of the treatment groups. The terms were chosen to set the power using the unstructured model at around 75%. Table III shows that if the 'true' covariance structure can be determined, then the resulting Wald statistics will be more powerful than if the unstructured model is adopted. This is perhaps unsurprising, but the reduction in power of up to 15% from adopting the unstructured model is non-negligible.

Inflation of the type 1 error rates using the Kenward-Roger adjustment is even more pronounced in design (C) where the actual test sizes found when using the unstructured covariance matrix are very far from the nominal level for the test of no treatment effect. However, it is noted again how the tests which rely on an estimate of the 'true' structure (where this has few parameters) appear to be reasonably robust, with actual sizes close to the nominal level. It is not appropriate to compare power here, as the actual test sizes have not all attained the nominal level.

To test the approach using an adjusted sandwich estimator we again consider simulations of study designs (A), (B) and (C). Results from 1000 simulations are shown in Table IV for the unadjusted χ^2 and F-Wald statistics using the sandwich estimator both with and without the bias adjustment proposed by Mancl and DeRouen. Also shown are the results obtained when the Pan and Wall (PW) adjusted statistic is adopted in each case. The working covariance matrix $\mathbf{W}$ is assumed to be the identity, so that ordinary least squares mean estimates are adopted.

Looking at the table it is clear in testing the (multiple parameter) hypotheses of no treatment/time interaction in the simple repeated measures studies of designs (A) and (B), and no treatment effect in the cross-over study of design (C) that adjusted tests using the Pan and Wall approach without bias adjustment are still somewhat inflated, but that the combined adjustment (accounting for both bias and variability) comes close to achieving nominal levels

Table IV. Summary of results from 1000 simulations of study designs (A)-(C): Wald tests using χ^2 and F distributions with an adjusted sandwich estimator.

Underlying 'true' covariance structure	Method of inference	Proportion of significant test results			
		Size			Power
		χ^2	F	PW	(PW)
Study design (A)					
Identity	Sandwich	0.426	0.388	0.115	—
	Bias-adj.	0.248	0.207	0.055	0.629
Compound Symmetry	Sandwich	0.419	0.394	0.119	—
	Bias-adj.	0.254	0.223	0.057	0.660
AR1 ($\rho = 0.2$)	Sandwich	0.428	0.391	0.142	—
	Bias-adj.	0.275	0.249	0.068	0.659
AR1 ($\rho = 0.8$)	Sandwich	0.416	0.383	0.166	—
	Bias-adj.	0.275	0.249	0.083	0.793
Antedependence	Sandwich	0.422	0.385	0.109	—
	Bias-adj.	0.257	0.225	0.047	0.593
Unstructured	Sandwich	0.423	0.390	0.109	—
	Bias-adj.	0.269	0.230	0.051	0.607
Unstr. (QRE)	Sandwich	0.430	0.386	0.206	—
	Bias-adj.	0.257	0.228	0.100	0.824
Study design (B)					
Identity	Sandwich	0.482	0.449	0.135	—
	Bias-adj.	0.293	0.252	0.056	0.569
Compound symmetry	Sandwich	0.444	0.396	0.130	—
	Bias-adj.	0.272	0.229	0.051	0.523
AR1 ($\rho = 0.2$)	Sandwich	0.485	0.444	0.152	—
	Bias-adj.	0.290	0.264	0.068	0.592
AR1 ($\rho = 0.8$)	Sandwich	0.452	0.399	0.160	—
	Bias-adj.	0.253	0.221	0.064	0.677
Antedependence	Sandwich	0.494	0.459	0.132	—
	Bias-adj.	0.292	0.259	0.055	0.514
Unstructured	Sandwich	0.459	0.411	0.088	—
	Bias-adj.	0.247	0.211	0.042	0.392
Unstr. (QRE)	Sandwich	0.397	0.349	0.136	—
	Bias-adj.	0.210	0.189	0.063	0.546
Study design (C)					
Identity	Sandwich	0.388	0.339	0.094	—
	Bias-adj.	0.225	0.196	0.039	
Compound symmetry	Sandwich	0.397	0.364	0.081	—
	Bias-adj.	0.225	0.198	0.034	
AR1 ($\rho = 0.2$)	Sandwich	0.384	0.345	0.098	—
	Bias-adj.	0.214	0.185	0.037	
AR1 ($\rho = 0.8$)	Sandwich	0.380	0.349	0.073	—
	Bias-adj.	0.215	0.188	0.034	
Antedependence	Sandwich	0.372	0.336	0.051	—
	Bias-adj.	0.208	0.178	0.018	
Unstructured	Sandwich	0.378	0.337	0.067	—
	Bias-adj.	0.209	0.184	0.025	
Unstr. (QRE)	Sandwich	0.335	0.294	0.027	—
	Bias-adj.	0.132	0.106	0.009	

Table gives proportion of type 1 errors (Size) and Power assessed using a fixed level of $\alpha = 0.05$. Power is calculated in Designs (A) and (B) by adding a linear effect in time to one treatment group to achieve approximately 75% using the Kenward-Roger adjustment assuming the given covariance model - see Table III. PW - Wald test using a sandwich estimator adjusted using the method of Pan and Wall with bias-adjustment as indicated. Power is not assessed where the test size is not fixed at the nominal level (—).

in each of these designs. It would appear that the Pan and Wall approach, of an adjusted F-test using the sandwich estimator combined with a suitable bias adjustment, leads to an appropriate test based on the general linear hypothesis for small sample repeated measures which are normally distributed.

The final column of Table IV shows comparable power levels from study designs (A) and (B) when adopting the Pan and Wall adjusted Wald tests using a bias adjusted sandwich estimator for the covariance estimator of the fixed effects. These are directly comparable with the Kenward-Roger adjusted Wald tests of Table III and show the relative poor performance of the adjusted sandwich estimator approach with regards power to detect departures from the null hypothesis, where both tests have fixed the size of the test. No comparison is possible for study design (C), although it should be noted that the bias-adjusted sandwich estimator has resulted in a test with nominal properties for this setting whereas the Kenward-Roger adjusted Wald test has not.

5. Discussion

The simulation studies of the previous section show that, in general, as the sample size decreases or the model becomes complex, either by design or due to missing values:

(1) REML based likelihood ratio tests of goodness of fit (and AIC_R/BIC_R) increasingly favour the unstructured covariance model in preference to the 'true' structure, so are not necessarily appropriate means of determining fit.

(2) The power of (small sample adjusted) tests of fixed effects using the unstructured covariance estimate is diminished in detecting departures from the null hypothesis.

(3) Where an exact small sample test of the fixed effects using the unstructured model is not possible, the approximate test is unlikely to have nominal properties. That is the actual size of the test may not be fixed at the nominal level, making inferences uncertain.

These simulations point to a dilemma in the use of likelihood based methods in such small samples. We have seen that where the covariance structure plays a part in the estimation of the fixed effects, poor nominal behaviour of inferences will follow unless there exists a parsimonious covariance structure that can be identified. However, the very small sample means that such identification is difficult if not impossible from the data alone. Simply guessing a parsimonious structure to use is not an appropriate solution as any lack of fit will have potentially serious consequences for validity of the subsequent inferences. This raises questions about the role of the covariance structure in making inferences in a repeated measures context where the sample size is very small and there is no external justification for a simple covariance structure.

We have seen that attempts to circumvent this problem through the use of GEE's with a small sample adjusted empirical sandwich estimator and Wald test can be successful in achieving a test size of nominal level. However, this test lacks power relative to the Kenward-Roger adjusted test, where comparisons are possible. The main problem with this approach is again due to the imprecision of the estimated covariance matrix of the data. Although eliminated from the estimation step, this is still central to the estimation of precision, and it can be observed from the simulation results (not shown) that the distribution of the resulting sandwich estimates of variance, while located in the right place, have very large dispersion. This in turn affects the power of the tests (and length of confidence intervals) that are derived from these variance estimates. A natural extension to this approach then is to consider removing the estimated covariance structure both from the estimates of precision and from the estimates of variability. That is, to develop a method which ignores the covariance structure completely when calculating the test statistic, using ordinary regression or ANOVA, and then corrects the null distribution for the observed dependency in the repeated measurements. This will be the focus of the following companion paper, where further consideration of the performance of the adjusted sandwich estimator will also be presented.

Appendix A

The following (5×5) symmetric matrices are used as the underlying covariance structures for the data generated in the simulations of Designs (A), (B) and (C) reported in Section 4.

A.1. Identity

$$\sigma^2 \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

with $\sigma^2 = 1$.

A.2 Compound Symmetry

$$\begin{bmatrix} \sigma_1^2+\sigma^2 & & & & \\ \sigma_1^2 & \sigma_1^2+\sigma^2 & & & \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2+\sigma^2 & & \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2+\sigma^2 & \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2+\sigma^2 \end{bmatrix} = \begin{bmatrix} 2 & & & & \\ 1 & 2 & & & \\ 1 & 1 & 2 & & \\ 1 & 1 & 1 & 2 & \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix}$$

with $\sigma_1^2 = 1$ and $\sigma^2 = 1$. Equivalently, the Compound Symmetry form can be considered as the 'Uniform Correlation' model, with $\sigma^2 = 2$ and $\rho = 0.5$,

$$\sigma^2 \begin{bmatrix} 1 & & & & \\ \rho & 1 & & & \\ \rho & \rho & 1 & & \\ \rho & \rho & \rho & 1 & \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}$$

A.3. AR1

$$\sigma^2 \begin{bmatrix} 1 & & & & \\ \rho & 1 & & & \\ \rho^2 & \rho & 1 & & \\ \rho^3 & \rho^2 & \rho & 1 & \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Two forms of AR1 are used, both with $\sigma^2 = 1$, and with $\rho = 0.2$ and $\rho = 0.8$ respectively.

$$\begin{bmatrix} 1 & & & & \\ 0.2 & 1 & & & \\ 0.04 & 0.2 & 1 & & \\ 0.008 & 0.04 & 0.2 & 1 & \\ 0.0016 & 0.008 & 0.04 & 0.2 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & & & & \\ 0.8 & 1 & & & \\ 0.64 & 0.8 & 1 & & \\ 0.512 & 0.64 & 0.8 & 1 & \\ 0.4096 & 0.512 & 0.64 & 0.8 & 1 \end{bmatrix}$$

A.4. Antedependence

$$\begin{bmatrix} \sigma_1^2 & & & & \\ \sigma_1\sigma_2\rho_1 & \sigma_2^2 & & & \\ \sigma_1\sigma_3\rho_1\rho_2 & \sigma_2\sigma_3\rho_2 & \sigma_3^2 & & \\ \sigma_1\sigma_4\rho_1\rho_2\rho_3 & \sigma_2\sigma_4\rho_2\rho_3 & \sigma_3\sigma_4\rho_3 & \sigma_4^2 & \\ \sigma_1\sigma_5\rho_1\rho_2\rho_3\rho_4 & \sigma_2\sigma_5\rho_2\rho_3\rho_4 & \sigma_3\sigma_5\rho_3\rho_4 & \sigma_4\sigma_5\rho_4 & \sigma_5^2 \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ 1.6 & 4 & & & \\ 1.44 & 3.6 & 9 & & \\ 0.768 & 1.92 & 4.8 & 16 & \\ 0.192 & 0.48 & 1.2 & 4 & 25 \end{bmatrix}$$

i.e. with $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $\sigma_3^2 = 9$, $\sigma_4^2 = 16$, $\sigma_5^2 = 25$, $\rho_1 = 0.8$, $\rho_2 = 0.6$, $\rho_3 = 0.4$, and $\rho_4 = 0.2$.

A.5. Unstructured

$$\begin{bmatrix} \sigma_1^2 & & & & \\ \sigma_{21} & \sigma_2^2 & & & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ 0.28 & 2 & & & \\ 1.04 & 1.96 & 3 & & \\ 0.80 & 0.56 & 2.08 & 4 & \\ 1.79 & 1.89 & 3.10 & 1.79 & 5 \end{bmatrix}$$

which has variance/correlation matrix,(variances on the diagonal, correlations below)

$$\begin{bmatrix} 1 & & & & \\ 0.2 & 2 & & & \\ 0.6 & 0.8 & 3 & & \\ 0.4 & 0.2 & 0.6 & 4 & \\ 0.8 & 0.6 & 0.8 & 0.4 & 5 \end{bmatrix}$$

A.6. Unstructured (QRE)

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & & \\ \sigma_{21} & \sigma_2^2 & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}^T + \sigma^2 \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & & & & \\ 3 & 10 & & & \\ 7 & 21 & 50 & & \\ 13 & 39 & 91 & 170 & \\ 21 & 63 & 147 & 273 & 442 \end{bmatrix}$$

i.e. with $\sigma^2 = 1$ and $\sigma_1^2 = \sigma_{21} = \sigma_2^2 = \sigma_{31} = \sigma_{32} = \sigma_3^2 = 1$. This has variance/correlation matrix

$$\begin{bmatrix} 2 & & & & \\ 0.67 & 10 & & & \\ 0.70 & 0.94 & 50 & & \\ 0.71 & 0.95 & 0.99 & 170 & \\ 0.71 & 0.95 & 0.99 & 1.00 & 442 \end{bmatrix}$$

References

1. Skene SS, Kenward MG. The analysis of very small samples of repeated measurements II: A modified Box correction. *Statistics in Medicine*. DOI: 10.1002/sim4072.

2. Brammer RJ. Modelling covariance structure in ascending dose studies of isolated tissues and organs. *Pharmaceutical Statistics* 2003; **2**:103–112. DOI: 10.1002/pst.30.

3. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. New York: Springer, 2000.

4. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine* 2002; **21**:1429-1441. DOI: 10.1002/sim.1142.

5. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2002; **57**:126-134. DOI: 10.1111/j.0006-341X.2001.00126.x.

6. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. *Biometrika* 1971; **58**:545-554. DOI: 10.1093/biomet/58.3.545.

7. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. Chichester: Wiley, 2007.

8. Kenward MG. A method for comparing profiles of repeated measurements. *Applied Statistics* 1987; **36**:296–308. DOI: 10.2307/2347788.

9. Wolfinger RD. Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological and Environmental Statistics* 1996; **1**:205–230. DOI: 10.2307/1400366.

10. Diggle PJ. An approach to the analysis of repeated measures. *Biometrics* 1988; **44**:959–971. DOI: 10.2307/2531727.

11. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19**:716–723. DOI: 10.1109/TAC.1974.1100705.

12. Schwarz G. Estimating the dimension of a model. *Annals of Statistics* 1978; **6**:461–464. DOI: 10.1214/aos/1176344136.

13. Longford NT. *Random coefficient models* Oxford University Press, 1990.

14. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**:983–997. DOI: 10.2307/2533558.

15. Kenward MG, Roger JH. An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis*, 2009; **53**:2583–2595. DOI: 10.1016/j.csda.2008.12.013.

16. Satterthwaite FE. An Approximate Distribution of estimates of variance components. *Biometrics Bulletin* 1946; **2**:110-114. DOI: 10.2307/3002019.

17. Schaalje GB, McBride JB, Fellingham GW. Adequacy of approximations to distributions of test statistics in complex linear models. *Journal of Agricultural, Biological and Environmental Services* 2002; **7**:512–524. DOI: 10.1198/108571102726.

18. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22. DOI: 10.1093/biomet/73.1.13.

19. Brown HK, Prescott RJ. *Applied Mixed Models in Medicine* Chichester: Wiley , 1999.

20. Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data* Oxford University Press, 1994.

21. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 2001; **96**:1387–1396. DOI: 10.1198/016214501753382309.