UPSS

Working Paper Series

Estimation of Semiparametric Mixed Analyis of Covariance Model

Virgelio M. Alao, Joseph Ryan G. Lansangan and Erniel B. Barrios

NO. 2017-03

October 2017

ESTIMATION OF SEMIPARAMETRIC MIXED ANALYSIS OF COVARIANCE MODEL

Virgelio M. Alao
Visayas State University
Joseph Ryan G. Lansangan
Erniel B. Barrios
University of the Philippines Diliman

ABSTRACT

A semiparametric mixed analysis of covariance model is postulated. This model is estimated by imbedding restricted maximum likelihood estimation and smoothing splines regression into the backfitting algorithm. Bootstrap method is further incorporated into the algorithm. The heterogeneous effect of covariates across groups of experimental units is postulated to affect the response through a nonparametric function to mitigate overparameterization. Simulation studies exhibited the capability of the postulated model (and estimation procedures) in increasing predictive ability and stabilizing variance components estimates even for small sample size and with minimal covariate effect, and regardless of the extent of misspecification error.

Keywords: *mixed ANCOVA model, nonparametric regression, backfitting, bootstrap, random effects, variance components*

1. Introduction

Linear models have been very popular in explaining continuous response variables through some pre-determined predictors. Given a linear model, the method of ordinary least squares (OLS) provides optimal estimates of the regression coefficients (Arnold, 1981). However, optimality of OLS is ensured only when the model is correctly specified. There are many situations where the linear model is too stringent to fit to the data (Christensen, 2011; Keppel and Winkens, 2004; Wu, 2010). There are also instances where some predictors are erroneously measured (Sheather, 2009). It is also possible that the measured proximate indicator for a predictor is a very tentative representation of the target predictor. Many solutions are proposed to resolve these issues (Dornheim and Brazauskas, 2011; Huitema, 1980; Milliken and Johnson, 2009), among them, the use of a nonlinear regression model, where the link function, i.e., the functional form on the dependence of the response on one or more of the predictors, is assumed to follow a nonlinear structure. In a more general scenario, the response is not assumed to follow a parametric link with the predictors, but rather a nonparametric function (Hardle, 1994; Mooney and Duval, 1993; Wang, 2011).

In experimental studies, one is interested in estimating the effect of a factor to a response. Thus, experimental units are carefully chosen so that all other factors that could affect the response are controlled (Montgomery, 2013). In the course of the experimentation, all conditions that might influence the response are controlled so that the "effect" manifested by experimental units can be attributed solely to the treatment that is varied among group of experimental units. In many cases though, the only experimental units available are not really homogeneous. In the same context, it is also possible that certain factors cannot be controlled during experimentation. For example, in *In Vivo* experiments, the diet of the subjects cannot

be controlled. In this case, the experimenter would measure variables/indicators that could differentiate experimental units (accounting for heterogeneity at the start of the study) or measure the factors that were not controlled during experimentation. These covariates are then included in the analysis.

In an experiment that includes some covariates like those described above, analysis of covariance is used to measure the treatment effect which is the main goal of the experiment. The covariates could exert influence on the response variable and thus cannot be ignored in the analysis. However, the effect of covariates is "set aside" to facilitate correct measurement of the treatment effect (Montgomery, 2013; Searle, Casella and McCulloch, 1992). And oftentimes, these covariates are assumed to have linear effect on the response variable.

In this paper, we propose to account for the effect of the covariate to follow a nonparametric structure. This will ensure that the portion of the response attributed to the covariate is appropriately accounted for and set aside so that the treatment effect can be correctly measured. A semiparametric model with nonparametric covariate effects and the random treatment effects assumed to follow a parametric structure is thus postulated.

This study aims (1) to postulate a mixed ANCOVA model in a semiparametric framework, (2) to estimate the postulated model through a hybrid backfitting algorithm, and (3) to evaluate the proposed procedures relative to the standard procedure in terms of predictive ability and stability of variance components estimates through simulation.

2. Parametric Mixed ANCOVA Model

The parametric mixed analysis of covariance (ANCOVA) model where the variable of interest, Y, is represented as a function of a covariate and the factors. This is given by (Montgomery, 2013; Keppel and Winkens, 2004)

$$Y_{ijk} = X_{ijk}\beta + \tau_j + \delta_k + (\tau\delta)_{jk} + \varepsilon_{ijk} \qquad \begin{cases} i = 1, 2, \dots, n\ observations \\ j = 1, 2, \dots, p\ treatments \\ k = 1, 2, \dots, q\ treatements \end{cases} \tag{1}$$

where

Y_{ijk} is the response variable score of the ijk^{th} observation
X_{ijk} is fixed covariate score of the ijk^{th} observation
β is the linear regression coefficient Y on X
τ_j, δ_k are random effects of treatment j and treatment k, respectively
$(\tau\delta)_{jk}$is the random interaction effect of the jk^{th} combination
ε_{ijk} is the error component such that $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$

By structure, the model (1) assumes the following (Arnold, 1981; Christensen, 2011; Keppel and Winkens, 2004; Montgomery, 2013;):

1. Randomization: randomly selecting subjects from some defined population and randomly and independently assigning subjects to treatment groups.
2. Homogeneity of within-group regressions.
3. Statistical independence of covariate and treatment.
4. Fixed covariate values that are error free.
5. Linearity of within-group regressions.
6. Normality of conditional Y scores.
7. Homogeneity of variance of conditional Y scores.
8. Random treatment levels.

This model, however, does not depict the presence of atypical observations, e.g. due to the occurrence of overdispersion or heterogeneity caused by the covariate (Wu, 2010; Dornheim and Brazauskas, 2011). When there is overdispersion or heterogeneity, some of the very important assumptions are not met such as assumptions 2 and 5 to 7 above. This problem has

substantial consequences to the bias of the ANCOVA F-statistic as cited by Huitema (1980), i.e., the bias introduced into the analysis by nonnormality of the dependent variable is greater when X is not normally distributed. In lieu of this problem, to account the varying effect of covariate across treatment-level factors, its effect on the dependent variable was postulated in a nonparametric framework.

3. Postulated Semiparametric Mixed ANCOVA Model

A semiparametric mixed ANCOVA model is postulated with nonparametric fixed covariate and parametric random effects. This is given by

$$Y_{ijk} = f(X_{ijk}) + \tau_j + \delta_k + (\tau\delta)_{jk} + \varepsilon_{ijk} \qquad \begin{cases} i = 1, 2, \dots, n\ observations \\ j = 1, 2, \dots, p\ treatments \\ k = 1, 2, \dots, q\ treatements \end{cases} \tag{2}$$

where

Y_{ijk} is the response variable score of the ijk^{th} observation
X_{ijk} is fixed covariate score of the ijk^{th} observation
$f(X_{ijk})$ is the smoothed function of X_{ijk} (nonparametric)
τ_j, δ_k are random effects of treatment j and treatment k respectively
$(\tau\delta)_{jk}$ is the random interaction effect of the jk^{th} combination
ε_{ijk} is the error component such that $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$

The assumptions of the postulated model are as follows:

1. Y is a continuous response variable.
2. X is fixed and a continuous covariate. The smoothed function of X is used in lieu of varying coefficients of the covariate between treatment-level factors.
3. τ_j, δ_k are random effects. These components are used to address the effect of treatment levels.
4. $(\tau\delta)_{jk}$ is random interaction effect. This component is used to address the effect of jk^{th} combination levels.
5. The predictors are independent of each other.
6. Randomization: randomly selecting subjects from some defined population and randomly and independently assigning subjects to treatment groups.

4. Estimation Procedure

This paper proposes two estimation procedures. The first procedure is a modified, iterative estimation infusing the restricted maximum likelihood or REML (Corbeil and Searle, 1976; Milliken and Johnson, 2009; Montgomery, 2013) and nonparametric regression via smoothing splines (Siminoff, 1996; Wang, 2011) into a backfitting framework (Hastie and Tibshirani 1990; Wood, 2006). The second procedure incorporates a bootstrap approach (Davison and Hinkley, 1997; Mooney and Duval, 1993) on the first procedure. The performance of these proposed procedures is evaluated in the postulated model (2) through simulated data. From this point onwards, the first procedure is named as ANCOVA via REML with splines or ARMS, and the second procedure as Bootstrap ANCOVA via REML with splines or B-ARMS.

The main idea of the ARMS procedure is to alternately estimate the parametric part corresponding to the variance components for the random-effects model and the nonparametric part corresponding to the smooth function of the covariate. The method can mitigate contamination where the conventional REML may possibly become problematic because of its assumptions of normality. The variance components are first estimated using the REML by ignoring the effect of the nonparametric component. The smooth function of covariate is then estimated via nonparametric regression using the residuals computed after fitting the model

with only the estimated parametric components. In the B-ARMS approach, the estimates obtained in ARMS are used and resampling with replacement of the residuals is applied. Details are discussed in the succeeding sections.

Algorithm of ARMS

Step1. Using the REML method, fit the parametric part of the model by ignoring $f(X_{ijk})$ so that the random-effects model would be

$$Y_{ijk} = \tau_j + \delta_k + (\tau\delta)_{jk} + \varepsilon_{ijk}$$

This contains the estimates of the variance components. We are of course much interested in the estimation of variance components over the means under a random-effects model.

Step2. Compute the partial residuals $e_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$. The partial residuals contain information on $f(X_{ijk})$ and thus will be used to estimate the $f(X_{ijk})$.

Step3. Estimate $f(X_{ijk})$ nonparametrically using smoothing spline.

Step4. Compute the new partial residuals $e^*_{ijk} = Y_{ijk} - \hat{f}(X_{ijk})$. The new partial residuals contain information on the parametric part and thus will be used to estimate the variance components in Step 1.

Step5. Repeat Step1 to Step4 until $\hat{f}(X_{ijk})$ and the variance components estimates do not change more than the tolerance level (say, 0.001).

Algorithm of B–ARMS

Step1. Obtain the initial estimates and residuals $\phi_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$by fitting the model (2) using ARMS. The residuals are used to obtain *R* bootstrap samples of residuals in Step2 below while the estimates are used to compute the new values of dependent variable in Step3 below.

Step2. Sample ϕ_{ijk} (Step1) with replacement from {1, 2, …, n} to have new set of residuals ϕ^*_{ijk}. The residuals ϕ^*_{ijk} are used to compute the new values of Y in Step3.

Step3. Compute the new values of dependent variable by $Y^*_{ijk} = \hat{f}(X_{ijk}) + \hat{\tau}_j + \hat{\delta}_k + \widehat{(\tau\delta)}_{jk} + \phi^*_{ijk}$ where $\hat{f}(X_{ijk})$, $\hat{\tau}_j$, $\hat{\delta}_k$ and $\widehat{(\tau\delta)}_{jk}$ are estimates from Step1 and new residuals ϕ^*_{ijk} from Step2.

Step4. Fit the model from the pseudo data in Step3 using ARMS. This contains the estimates of the variance components from the dependent variable Y^*_{ijk}.

Step5. Repeat Step2 to Step4 *R* times. In this paper, $R = 200$.

5. Simulation Study

The postulated model in this study with the estimation procedures will be evaluated using simulated data for balanced design. Each data set was composed of n_{jk} treatment combination size/replicate, j = 1, …, p treatments and k = 1, …, q treatments. Y_{ijk} was computed as a function of a covariate and the group-level factors following $Y_{ijk} = f(X_{ijk}) + m\tau_j + m\delta_k + (\tau\delta)_{jk} + w\varepsilon_{ijk}$ where m is a constant value used to determine the minimal or dominating effect of the covariate/factors. When the value of m is small, the effect of covariate is dominating, i.e. approximately 75% of the variability in Y is explained by the covariate. When the value of m is large, the effect of covariate is minimal, i.e. approximately 20% of the variability in Y is explained by the covariate. The variability in Y explained by the covariate X is determined using a structure coefficient squared or Pearson correlation squared ($r^2_{X,Y}$) between X and Y (Nathans et al. (2012)).

The constant w in $Y_{ijk} = f(X_{ijk}) + m\tau_j + m\delta_k + (\tau\delta)_{jk} + w\varepsilon_{ijk}$ is used to induce the misspecification error. There is much misspecification in the model when the assigned value of w is large. Also, the $f(X_{ijk})$ has two forms: linear and nonlinear functions. The linear function $f(X_{ijk}) = \beta X_{ijk}$ is generated from the normal distribution and subjected for contamination (5% and 10% of the treatment combinations). For the covariate contamination, a change in the value of β is applied. The nonlinear function $f(X_{ijk}) = \exp(\beta X_{ijk})$ is generated from the exponential family and introduced in the model to manifest heterogeneity/overdispersion and to further verify the predictive ability of the proposed methods in the presence of nonlinearity/heterogeneity. Under this setting, the values of m as well as the variances of the factors were changed accordingly to satisfy the minimal/dominating effect of the covariate in the model.

The treatment effects are assumed to be normally distributed. Table 1 summarizes the simulation settings.

Table1. Boundaries of Simulation Study

1. distribution of X_{ijk}	$X_{ijk} \sim N(20, 2^2)$
2. functional form of $f(X_{ijk})$	βX_{ijk} $\exp(\beta X_{ijk})$
3. value of β	$\beta = 1; \beta_c = 0.05, 2$ (contamination)
4. distribution of τ_j	$\tau_j \sim N(0, 1.5^2) * m$
5. distribution of δ_k	$\delta_k \sim N(0, 1.5^2) * m$
6. distribution of $(\tau\delta)_{jk}$	$(\tau\delta)_{jk} \sim N(0, 0.5^2)$
7. constant m to adjust the contribution of the covariate/factors on Y	m = 0.5 and 1.7
8. distribution of ε_{ijk}	$\varepsilon_{ijk} \sim N(0,1)$
9. number of treatments: p, q	small size – 2 levels for each factor medium size – 10 levels for each factor large size – 20 levels for each factor
10. treatment combination/replicate size: n	small size – 2 medium size – 5 large size – 10
11. misspecification error w in the model	w = 1, 5

The simulated data was used to compare the proposed estimation procedures under the semiparametric mixed ANCOVA model to the conventional method, REML, through their mean absolute percentage error (MAPE), standard error of the variance components estimates (SE) and bias of the variance components estimates (BV or Bias). The MAPE tells the average percentage of the forecast errors to the actual values. The SE is the standard deviation of the variance components estimates. The closer the value of SE to zero, the more precise the estimate. When the BV is zero, the estimator is said to be unbiased. The method with smaller MAPE, SE and BV is considered better relative to other methods. The formulas are as follows:

$$MAPE = \frac{1}{npq} \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{q} \left| \frac{Y_{ijk} - \hat{Y}_{ijk}}{Y_{ijk}} \right| * 100$$

$$SE = \sqrt{\frac{1}{R-1}\sum_{h=1}^{R}(V_h - \bar{V})^2}$$

$$\text{BV} = \text{Bias} = \overline{V} - \theta$$

where

npq is the total number of data points

R is the total number of V_h observations

V_h is the h^{th} variance of the specific component, and $\overline{V}$ is the mean of $V_h s$

θ is the true variance of the specific component

6. Results and Discussion

The predictive performance of the proposed semiparametric estimation procedures – namely the ANCOVA via REML with smoothing splines (ARMS) and the bootstrap ANCOVA via REML with smoothing splines (B-ARMS) – are compared with the conventional REML in terms of the MAPE. Also, the bias and standard error of the estimates of the variance components were compared to further evaluate the viability of the estimation procedures. Smaller values of MAPE, standard error and bias would indicate better estimates and better predictive ability.

6.1 Contamination-free Case

The semiparametric model (2) is first assumed to have no heterogeneity and/or overdispersion to verify the performance of the proposed procedures under such ideal scenario.

6.1.1 Effect of the degree of contribution of covariate

The values of m reflect the relative contribution of a covariate in the model. The covariate is dominating if the value of m is small, i.e. approximately 75% of the variability in Y is explained by the covariate, while minimal role if the value of m is large, i.e. approximately 20% of the variability in Y is explained by the covariate. As shown in Table 2, when there is a dominating covariate ($m = 0.5$), MAPE is lower compared to when there is a minimal covariate ($m = 1.7$) in the model. The magnitude of bias decreases as the degree of contribution of covariate increases. The proposed methods have biased estimates potentially because the REML method is incorporated in the procedures, as discussed by Searle et al. (2006), REML estimators are biased. Though the proposed methods provide biased estimates, the predictive ability is comparably better than REML.

Table 2. Average MAPE, SE and Bias by covariate contribution for contamination-free scenario

Scenarios		**Bias($\hat{\sigma}_\tau^2$)**	**SE($\hat{\sigma}_\tau^2$)**	**Bias($\hat{\sigma}_\delta^2$)**	**SE($\hat{\sigma}_\delta^2$)**	**Bias($\hat{\sigma}_{\tau\delta}^2$)**	**SE($\hat{\sigma}_{\tau\delta}^2$)**	**MAPE**
ARMS	m = 0.5	-0.464	0.223	-0.472	0.227	-0.145	0.216	4.867
	m = 1.7	-5.942	1.347	-5.848	1.817	0.216	1.043	12.700
B-ARMS	m = 0.5	-0.416	0.209	-0.412	0.205	-0.113	0.185	4.367
	m = 1.7	-5.745	1.110	-5.570	1.176	0.338	0.940	11.267
REML	m = 0.5	-0.391	1.270	-0.490	0.166	-0.173	0.176	5.833
	m = 1.7	-5.938	1.240	-5.927	1.594	0.156	1.053	15.200

6.1.2 Effect of the size of experiment

The experiment size is categorized into three, namely, small (2x2x2), medium (10x10x5) and large (20x20x10). Simulations (see Table 3 below) show slightly lower values

of MAPE for small experiments compared to large experiments for both proposed methods. ARMS and B-ARMS estimates are biased for all experiment sizes but still comparable to the classical REML. The magnitude of bias increases as the experiment size increases. However, the amount of standard error decreases as the experiment size increases. Overall, the proposed methods produce relatively better estimates and higher predictive ability for the different experiment sizes compared to REML.

Table 3. Average MAPE, SE and Bias by experiment size for contamination-free scenario

Scenarios		**Bias($\hat{\sigma}_\tau^2$)**	**SE($\hat{\sigma}_\tau^2$)**	**Bias($\hat{\sigma}_\delta^2$)**	**SE($\hat{\sigma}_\delta^2$)**	**Bias($\hat{\sigma}_{\tau\delta}^2$)**	**SE($\hat{\sigma}_{\tau\delta}^2$)**	**MAPE**
ARMS	Small	-2.255	2.044	-2.153	2.648	0.394	1.492	7.267
	Medium	-3.074	0.059	-3.075	0.056	-0.146	0.163	9.000
	Large	-3.100	0.009	-3.100	0.011	-0.231	0.030	9.000
B-ARMS	small	-1.925	1.729	-1.695	1.805	0.624	1.376	5.967
	medium	-3.068	0.050	-3.069	0.049	-0.163	0.125	8.233
	Large	-3.099	0.008	-3.099	0.008	-0.231	0.029	8.300
REML	small	-2.174	2.946	-2.275	2.298	0.298	1.489	8.500
	medium	-3.076	0.051	-3.078	0.046	-0.164	0.137	10.767
	Large	-3.101	0.008	-3.100	0.008	-0.234	0.027	10.933

6.1.3 Comparison of the proposed methods and REML

In general, the proposed estimation procedures exhibit reasonable results in the absence of heterogeneity/nonlinearity/outliers (see Table 4). The procedures produced considerably smaller MAPE than REML. The B-ARMS had the smallest value of MAPE and standard error compared to the others. In terms of bias, the three methods are at par.

Table 4. Average MAPE, SE and Bias by methods for contamination-free scenario

Method	**Bias($\hat{\sigma}_\tau^2$)**	**SE($\hat{\sigma}_\tau^2$)**	**Bias($\hat{\sigma}_\delta^2$)**	**SE($\hat{\sigma}_\delta^2$)**	**Bias($\hat{\sigma}_{\tau\delta}^2$)**	**SE($\hat{\sigma}_{\tau\delta}^2$)**	**MAPE**
ARMS	-2.809	0.704	-2.776	0.905	0.006	0.562	8.422
B-ARMS	-2.697	0.595	-2.621	0.621	0.077	0.510	7.500
REML	-2.784	1.002	-2.818	0.784	-0.033	0.551	10.067

6.2 Contamination Case

In this scenario, contamination in the covariate was incorporated to evaluate the performance of the proposed procedures. Two levels of contamination (5% and 10%) corresponding to two values of β_c (0.5 and 2) were considered.

6.2.1 Effect of the degree of contribution of covariate

As shown in Table 5, with $\beta_c = 0.05$, when there is a dominating covariate ($m = 0.5$), MAPE is slightly greater compared to when there is a minimal covariate ($m = 1.7$) in the model. This is true for both low level (5%) and high level (10%) of contaminations. Note however that as more observations are contaminated in the covariate, predictive ability deteriorates.

Table 5. Average MAPE, SE and Bias by covariate contribution for contamination scenario ($\beta_c = 0.05$)

Scenarios		**Bias($\hat{\sigma}^2_\tau$)**	**SE($\hat{\sigma}^2_\tau$)**	**Bias($\hat{\sigma}^2_\delta$)**	**SE($\hat{\sigma}^2_\delta$)**	**Bias($\hat{\sigma}^2_{\tau\delta}$)**	**SE($\hat{\sigma}^2_{\tau\delta}$)**	**MAPE**
5 %								
ARMS	m = 0.5	11.962	0.373	-0.528	0.054	4.063	0.290	24.150
	m = 1.7	5.189	0.838	-6.406	0.154	3.392	0.784	15.850
B-ARMS	m = 0.5	12.228	0.182	-0.523	0.026	4.272	0.136	17.600
	m = 1.7	6.209	0.430	-6.397	0.069	4.079	0.350	15.500
REML	m = 0.5	-0.557	0.010	-0.558	0.008	-0.233	0.029	33.850
	m = 1.7	-6.471	0.055	-6.473	0.051	-0.153	0.153	26.900
10 %								
ARMS	m = 0.5	22.902	0.800	-0.462	0.236	-0.133	0.318	35.100
	m = 1.7	17.373	2.546	-5.775	1.761	0.306	1.168	23.467
B-ARMS	m = 0.5	22.981	0.400	-0.417	0.190	-0.082	0.173	30.100
	m = 1.7	17.701	1.468	-5.615	1.102	0.424	0.873	21.767
REML	m = 0.5	-0.347	1.012	-0.450	0.441	-0.173	0.179	58.467
	m = 1.7	-5.154	4.934	-5.646	2.356	0.178	1.056	36.667

Moreover, the variance component estimates are biased for the three methods with different degrees of contribution of covariate at different levels of contamination. ARMS and B-ARMS $\hat{\sigma}^2_\tau$ estimates are more biased than REML $\hat{\sigma}^2_\tau$ estimates because contamination is contained within one-group level only for the factor τ. The methods ARMS and B-ARMS are based on the backfitting algorithm that estimate first the random effect in the model, and the error variance is estimated last, thus resulting to larger errors. Such observation is similar to those cited by Santos and Barrios (2012).

When $\beta_c = 2$, with dominating covariate, smaller MAPE is observed than when the covariate has minimal role in the model (see Table 6 for details). This is true for both 5% and 10% contamination in the data. ARMS and B-ARMS estimates for $\hat{\sigma}^2_\tau$ are still more biased than REML $\hat{\sigma}^2_\tau$ estimates, in the same context, due to the inherent backfitting algorithm.

Table 6. Average MAPE, SE and Bias by covariate contribution for contamination scenario ($\beta_c = 2$)

Scenarios		**Bias($\hat{\sigma}^2_\tau$)**	**SE($\hat{\sigma}^2_\tau$)**	**Bias($\hat{\sigma}^2_\delta$)**	**SE($\hat{\sigma}^2_\delta$)**	**Bias($\hat{\sigma}^2_{\tau\delta}$)**	**SE($\hat{\sigma}^2_{\tau\delta}$)**	**MAPE**
5 %								
ARMS	m = 0.5	7.222	0.606	-1.102	0.036	2.366	0.411	5.800
	m = 1.7	3.256	0.715	-3.769	0.096	1.068	0.618	15.100
B-ARMS	m = 0.5	7.656	0.212	-1.095	0.028	2.633	0.159	5.450
	m = 1.7	4.443	0.515	-3.741	0.081	1.550	0.410	14.000
REML	m = 0.5	-1.119	0.010	-1.120	0.008	-0.233	0.029	13.400
	m = 1.7	-3.793	0.055	-3.795	0.051	-0.153	0.153	34.200
10 %								
ARMS	m = 0.5	15.232	1.270	-1.027	0.225	-0.148	0.271	8.533
	m = 1.7	12.687	2.774	-3.102	1.761	0.280	1.145	18.867
B-ARMS	m = 0.5	15.533	0.467	-0.982	0.199	-0.109	0.176	8.567
	m = 1.7	13.334	1.704	-2.931	1.127	0.409	0.926	21.267
REML	m = 0.5	-0.903	1.085	-1.008	0.445	-0.174	0.178	18.633
	m = 1.7	-2.503	5.379	-3.006	2.383	0.180	1.069	43.633

6.2.2 Effect of experiment size

The two proposed methods show lower values of MAPE for small experiment size than the REML as shown in Table 7 and Table 8. This is true for $\beta_c = 0.05$ and $\beta_c = 2$. ARMS and B-ARMS estimates are biased for all experiment size but still comparable to the classical REML except for $\hat{\sigma}_\tau^2$ component. Again, this is because contamination is contained within one-group level only for the factor τ and the ARMS and B-ARMS procedures estimate first the random effect in the model. However, the amount of standard error decreases as the experiment size increases.

Table 7. Average MAPE, SE and Bias by experiment size for contamination case ($\beta_c = 0.05$)

Scenarios		**Bias($\hat{\sigma}_\tau^2$)**	**SE($\hat{\sigma}_\tau^2$)**	**Bias($\hat{\sigma}_\delta^2$)**	**SE($\hat{\sigma}_\delta^2$)**	**Bias($\hat{\sigma}_{\tau\delta}^2$)**	**SE($\hat{\sigma}_{\tau\delta}^2$)**	**MAPE**
5 %								
ARMS	medium	3.221	0.798	-2.989	0.184	7.689	0.979	18.633
	large	14.837	0.377	-3.098	0.013	-0.215	0.050	24.067
B-ARMS	medium	4.401	0.412	-2.974	0.087	8.589	0.456	15.867
	large	14.890	0.187	-3.099	0.005	-0.224	0.019	18.900
REML	medium	-3.076	0.051	-3.079	0.046	-0.164	0.138	37.600
	large	-3.100	0.008	-3.100	0.008	-0.234	0.027	35.733
10 %								
ARMS	small	-2.033	2.613	-2.056	2.551	0.467	1.779	17.700
	medium	32.664	1.530	-3.064	0.084	-0.098	0.218	43.867
	large	30.973	0.536	-3.098	0.012	-0.215	0.048	40.600
B-ARMS	small	-1.700	1.680	-1.762	1.709	0.749	1.325	13.300
	medium	32.854	0.702	-3.070	0.028	-0.130	0.073	41.433
	large	31.010	0.221	-3.100	0.004	-0.227	0.014	38.700
REML	small	-1.023	8.136	-1.832	3.923	0.324	1.505	61.167
	medium	-3.076	0.052	-3.079	0.045	-0.164	0.137	57.000
	large	-3.101	0.008	-3.100	0.008	-0.234	0.027	62.000

Table 8. Average MAPE, SE and Bias by experiment size for contamination case ($\beta_c = 2$)

Scenarios		**Bias($\hat{\sigma}_\tau^2$)**	**SE($\hat{\sigma}_\tau^2$)**	**Bias($\hat{\sigma}_\delta^2$)**	**SE($\hat{\sigma}_\delta^2$)**	**Bias($\hat{\sigma}_{\tau\delta}^2$)**	**SE($\hat{\sigma}_{\tau\delta}^2$)**	**MAPE**
5 %								
ARMS	medium	0.451	0.735	-2.332	0.112	3.454	1.001	9.467
	large	9.242	0.687	-2.394	0.011	-0.228	0.033	10.333
B- ARMS	medium	1.849	0.448	-2.299	0.094	4.263	0.511	8.567
	large	9.746	0.252	-2.394	0.008	-0.229	0.030	9.800
REML	medium	-2.371	0.051	-2.374	0.046	-0.164	0.138	21.433
	large	-2.395	0.008	-2.395	0.008	-0.234	0.027	22.267
10 %								
ARMS	small	-1.359	2.515	-1.344	2.566	0.465	1.747	9.233
	medium	20.617	2.731	-2.367	0.058	-0.132	0.178	14.233
	large	19.748	0.898	-2.394	0.012	-0.227	0.034	14.000
B-ARMS	small	-1.036	1.696	-1.051	1.738	0.720	1.336	7.600
	medium	21.759	1.006	-2.363	0.050	-0.151	0.128	14.367
	large	20.743	0.352	-2.393	0.008	-0.230	0.029	17.667
REML	small	-0.300	9.026	-1.201	3.841	0.313	1.482	15.967
	medium	-2.371	0.052	-2.374	0.046	-0.164	0.137	30.600
	large	-2.396	0.008	-2.395	0.008	-0.234	0.027	36.433

6.2.3 Comparison of the proposed methods and REML

Generally, looking at the result of the contamination of covariate either in low or large values of β_c, between the two proposed methods, B-ARMS has a good predictive ability and stability of the estimates as manifested in the lower MAPE and standard error, respectively (see Table 9). ARMS and B-ARMS have lower value of MAPE compared to REML.

Table 9. Average MAPE, SE and Bias by methods for contamination scenario

Method	Bias($\hat{\sigma}_\tau^2$)	SE($\hat{\sigma}_\tau^2$)	Bias($\hat{\sigma}_\delta^2$)	SE($\hat{\sigma}_\delta^2$)	Bias($\hat{\sigma}_{\tau\delta}^2$)	SE($\hat{\sigma}_{\tau\delta}^2$)	MAPE
$\beta_c = 0.05$							
ARMS	14.782	1.074	-2.891	0.490	1.894	0.598	27.703
B-ARMS	15.184	0.583	-2.840	0.313	2.157	0.354	24.264
REML	-2.744	1.381	-2.880	0.676	-0.112	0.319	48.361
$\beta_c = 2$							
ARMS	8.924	1.379	-2.199	0.470	0.824	0.585	11.194
B-ARMS	9.810	0.684	-2.141	0.325	1.065	0.384	11.197
REML	-2.036	1.529	-2.187	0.662	-0.113	0.316	24.758

Meanwhile, the two proposed methods outperform the REML method in terms of predictive ability. Also, the magnitude of bias of the estimates of variance components for the two methods proposed are comparable except for $\hat{\sigma}_\tau^2$ component. Again, this is because contamination is contained within one-group level only for the factor τ and the ARMS and B-ARMS procedures estimate first the random effect in the model.

6.3 Model Misspecification

Misspecification was applied in the model particularly in the error term to further investigate the performance of the proposed procedures with the postulated model. This was done by multiplying a considerable large value, say 5, to the error term to exhibit much variance of the residuals.

6.3.1 Effect of the degree of contribution of covariate

When there is a dominating covariate ($m = 0.5$), MAPE values are lower compared to when there is a minimal covariate ($m = 1.7$) in the model. The three methods produced biased estimates for the different degrees of contribution of covariate. The magnitude of bias decreases as the degree of contribution of covariate increases.

Table 10. Average MAPE, SE and Bias by covariate contribution for misspecification case

Scenarios		Bias($\hat{\sigma}_\tau^2$)	SE($\hat{\sigma}_\tau^2$)	Bias($\hat{\sigma}_\delta^2$)	SE($\hat{\sigma}_\delta^2$)	Bias($\hat{\sigma}_{\tau\delta}^2$)	SE($\hat{\sigma}_{\tau\delta}^2$)	MAPE
ARMS	m = 0.5	0.371	1.946	0.286	1.699	1.046	2.595	19.600
	m = 1.7	-5.060	3.026	-5.207	3.034	1.221	2.855	35.867
B-ARMS	m = 0.5	0.779	1.976	0.740	1.967	1.067	1.759	17.400
	m = 1.7	-4.526	3.006	-4.483	2.971	1.415	2.504	31.167
REML	m = 0.5	0.291	1.821	0.221	1.701	0.821	2.165	24.067
	m = 1.7	-5.165	3.009	-5.371	2.424	1.086	2.834	43.667

6.3.2 Effect of experiment size

The two proposed methods show smaller values of MAPE for small experiment sizes as shown in Table 11. Cases under the small experiment size show lower MAPE compared to

those under the large experiment size. ARMS and B-ARMS variance estimates are biased across all experiment sizes and are comparable to the REML. The magnitude of bias increases as the experiment size increases. However, the amount of standard error decreases as the experiment size increases. Hence, it may be inferred that the proposed methods produce reasonable estimates and higher predictive ability for the different experiment size even in a misspecified model.

Table 11. Average MAPE, SE and Bias by experiment size for misspecification scenario

Scenarios		**Bias($\hat{\sigma}_\tau^2$)**	**SE($\hat{\sigma}_\tau^2$)**	**Bias($\hat{\sigma}_\delta^2$)**	**SE($\hat{\sigma}_\delta^2$)**	**Bias($\hat{\sigma}_{\tau\delta}^2$)**	**SE($\hat{\sigma}_{\tau\delta}^2$)**	**MAPE**
ARMS	small	0.187	6.860	-0.149	6.545	3.348	7.167	18.800
	medium	-2.971	0.228	-2.996	0.194	0.155	0.693	28.200
	large	-3.081	0.041	-3.085	0.035	-0.177	0.133	30.700
B-ARMS	small	1.556	7.032	1.536	6.950	3.688	5.596	15.333
	medium	-2.955	0.209	-2.965	0.204	0.116	0.523	25.833
	large	-3.079	0.035	-3.081	0.033	-0.169	0.124	27.433
REML	small	-0.052	6.737	-0.455	5.787	2.896	6.559	22.300
	medium	-2.980	0.218	-3.007	0.169	0.083	0.583	34.033
	large	-3.083	0.036	-3.089	0.029	-0.185	0.116	38.667

6.3.3 Comparison of the proposed methods and REML

In general, the proposed estimation procedures are advantageous in increasing predictive ability and provide reasonable magnitude of bias and standard error in the presence of misspecification in the model (see Table 12). They have considerably smaller MAPE than the REML method. The B-ARMS has the smallest value of MAPE compared to other methods. ARMS, B-ARMS and REML estimates are at par in terms of the magnitude of bias.

Table 12. Average MAPE, SE and Bias by methods for misspecification scenario

Method	**Bias($\hat{\sigma}_\tau^2$)**	**SE($\hat{\sigma}_\tau^2$)**	**Bias($\hat{\sigma}_\delta^2$)**	**SE($\hat{\sigma}_\delta^2$)**	**Bias($\hat{\sigma}_{\tau\delta}^2$)**	**SE($\hat{\sigma}_{\tau\delta}^2$)**	**MAPE**
ARMS	-1.955	2.376	-2.076	2.258	1.109	2.664	25.900
B-ARMS	-1.493	2.425	-1.503	2.396	1.212	2.081	22.867
REML	-2.039	2.330	-2.184	1.995	0.931	2.419	31.667

6.4 Nonlinearity

This nonlinearity condition of the function $f(X_{ijk})$ is addressed to further evaluate the behavior of the proposed methods when the data is nonlinear and/or heterogeneous/overdispersed. To simulate nonlinearity, exponential function was introduced to the covariate.

6.4.1 Effect of the degree of contribution of covariate

The three methods under the dominating covariate ($m = 0.5$) show lower value of MAPE compared to with the minimal covariate ($m = 1.7$) in the model as shown in Table 13. The magnitude of bias decreases as the degree of contribution of covariate increases for the two proposed methods. The REML produced more biased estimates than the ARMS and B-ARMS for all degrees of contribution of the covariate. In addition, the REML produced higher values of MAPE than the ARMS and B-ARMS.

Table 13. Average MAPE, SE and Bias by covariate contribution for nonlinearity scenario

Scenarios		Bias($\hat{\sigma}_\tau^2$)	SE($\hat{\sigma}_\tau^2$)	Bias($\hat{\sigma}_\delta^2$)	SE($\hat{\sigma}_\delta^2$)	Bias($\hat{\sigma}_{\tau\delta}^2$)	SE($\hat{\sigma}_{\tau\delta}^2$)	MAPE
ARMS	m = 0.5	-203.44	46.09	-193.64	82.31	16.01	70.90	10.33
	m= 1.7	-2303.92	465.98	-2232.99	699.55	165.90	410.02	241.33
B-ARMS	m = 0.5	-189.70	48.64	-177.69	51.45	28.68	45.39	9.20
	m= 1.7	-2232.76	389.89	-2140.17	426.48	212.78	332.03	192.27
REML	m = 0.5	124041.78	123806.83	121462.88	123740.76	41.79	135.70	14.93
	m= 1.7	55698.64	58986.68	58692.99	77546.01	181.23	504.29	329.73

6.4.2 Effect of experiment size

The values of MAPE for the two methods proposed are smaller in all experiment sizes as shown in Table 14. Also, the three methods under the small experiment size have smaller values of MAPE and magnitude of bias compared to the medium and large experiment sizes. The REML produced more biased estimates than the ARMS and B-ARMS for most of the three experiment sizes. Furthermore, the REML produced higher values of MAPE than the ARMS and B-ARMS.

Table 14. Average MAPE, SE and Bias by experiment size for nonlinearity scenario

Scenarios		Bias($\hat{\sigma}_\tau^2$)	SE($\hat{\sigma}_\tau^2$)	Bias($\hat{\sigma}_\delta^2$)	SE($\hat{\sigma}_\delta^2$)	Bias($\hat{\sigma}_{\tau\delta}^2$)	SE($\hat{\sigma}_{\tau\delta}^2$)	MAPE
ARMS	small	-929.74	662.40	-821.33	1021	218.84	582.49	38.48
	medium	-1198	18.32	-1198	19.20	23.73	50.51	168.85
	large	-1206	2.97	-1206	3.71	-2.28	10.78	81.68
B-ARMS	small	-816.84	569.79	-675.54	622.43	305.59	458.10	30.35
	medium	-1196	16.61	-1196	16.62	19.33	41.43	125.18
	large	-1206	2.65	-1206	2.66	-2.51	9.76	78.68
REML	small	-839.08	888.94	-826.87	1069	285.61	812.86	60.94
	medium	115658	120132	115970	13650	23.65	51.67	226.59
	large	181067	182004	179531	183054	-1.65	11.11	109.07

6.4.3 Comparison of the proposed methods and REML

Among the three methods, B-ARMS showed the smallest value of MAPE (see Table 15). In general, when the nonlinearity/heterogeneity happens in the data caused by the covariate, the two proposed methods are preferable than the conventional REML. A big difference can be observed in the results of REML either in MAPE, bias or standard error from the two proposed procedures. This can be attributed to the fact that the REML procedure lacks the capacity to eliminate peculiar observations which distorts the results of the estimates and predictive ability in the presence of nonlinearity/heterogeneity/overdispersion.

Table 15. Average MAPE, SE and Bias by methods for nonlinearity scenario

Method	Bias($\hat{\sigma}_\tau^2$)	SE($\hat{\sigma}_\tau^2$)	Bias($\hat{\sigma}_\delta^2$)	SE($\hat{\sigma}_\delta^2$)	Bias($\hat{\sigma}_{\tau\delta}^2$)	SE($\hat{\sigma}_{\tau\delta}^2$)	MAPE
ARMS	-1111	228	-1075	348.11	80.10	214.59	96.34
B-ARMS	-1073	196	-1026	213.90	107.47	169.76	78.07
REML	98629	101008	98225	106876	102.53	291.88	132.20

7. Conclusions and Recommendations

A semiparametric mixed analysis of covariance model was postulated. The covariate effect follows a nonparametic function while keeping the parametric formulation for the random-effects component. The nonparametric function is considered as a mitigation strategy

for the possible contamination and nonlinearity of the covariates in the presence of heterogeneity/overdispersion in the data. We proposed two estimation procedures, (1) based on the imbedded restricted maximum likelihood and nonparametric regression in the backfitting algorithm, and (2) infusing bootstrap into the hybrid backfitting algorithm to increase the predictive ability and to enhance stability in estimates of the variance components, while relaxing the usual assumption of linearity/homogeneity of within-group regressions.

The simulation study confirmed the advantages of the two estimation procedures over the conventional restricted maximum likelihood. The proposed procedures yield comparable mean absolute percentage error (MAPE) and estimates when there is a good linear fit (contamination free) in the model. In cases where there are contaminations, advantages in predictive ability are observed in favor of the proposed methods. In general, the two methods are more favorable and advantageous (in terms of both estimation and predictive ability) than REML when linear model fit is poor (heterogeneous/overdispersed data). Small sized experiments and dominating covariates are enough prerequisite to achieve increase in predictive ability and stability of variance component estimates using the proposed methods over REML.

Further studies on the effect of unbalanced design may be considered and two or more covariates to investigate further the performance of these methods. Contamination from a more complicated distributions may also be considered. Contamination for every group-level may also be considered. Aside from exponential distribution, other distributions might also be considered for overdispersion.

REFERENCES

Arnold, S.F. (1981). The Theory of Linear Models and Multivariate Analysis. John Wiley and Sons, Inc. USA.

Christensen, R. (2011). Plane Answers to Complex Questions: The Theory of Linear Models. 4th ed. Springer Science+Business Media, LLC.

Corbeil, R.R. and Searle, S.R. (1976). Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. Technometrics, Vol 18, No. 1, pp 31-38.

Davison, A. C. and Hinkley, D. V. (1997). Cambridge Series in Statistical and Probabilistic Mathematics: Bootstrap Methods and Their Application. Cambridge University Press.

Dornheim, H. and Brazauskas, V. (2011) . Robust-efficient fitting of mixed linear models: Methodology and theory. Journal of Statistical Planning and Inference 141, 1422–1435

Hardle, W. (1994). Applied Nonparametric Regression. Springer.

Hastie, T. J. and Tibshirani R. J. (1990). Generalized Additive Models. Chapman and Hall/CRC, New York.

Huitema, B. E. (1980). The Analysis of Covariance and Alternatives. John Wiley & Sons, Inc.

Keppel, G. and Wickens, T. D. (2004). Design and Analysis: A Researcher's Handbook. 4th ed.

Milliken, G. A. and Johnson, D. E. (2009). Analysis of Messy Data: Vol.1, Designed Experiments. 2nd ed. Chapman & Hall/CRC, New York

Montgomery,D.C. (2013). Design and Analysis of Experiments. 8th ed. John Wiley & Sons, Inc.

Mooney, C. Z. and Duval, R. D. (1993). Bootstrapping: A Nonparametric Approach to Statistical Inference. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-095. New Burky Park, CA: Sage.

Santos, E. and Barrios, E. B. 2012. Nonparametric Decomposition of Time Series Data with Inputs. Communications in Statistics – Simulation and Computation, 41:9, 1693-1710.
Searle, S. R., Casella, G. and McCulloch, C. E. (1992). Variance Components. John Wiley and Sons, Inc.Hoboken, New Jersey.
Sheather, S. J. (2009). A Modern Approach to Regression with R. Springer Texts in Statistics.
Simonoff, J. S. (1996). Springer Series in Statistics: Smoothing Methods in Statistics. Springer-Verlag, N.Y. Inc.
Wang, Y. (2011). Monographs on Statistics and Applied Probability 121: Smoothing Splines: Methods and Applications. Chapman and Hall/CRC.
Wood, S. N. (2006). Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.
Wu, L. (2010). Monographs on Statistics and Applied Probability 113: Mixed Effects Models for Complex Data. Chapman and Hall/CRC.