Modeling High Dimensional Multilevel Data using the Lasso Estimator: A Simulation Study

W. Holmes Finch

Ball State University

Abstract

In some situations, researchers may be faced with the problem high dimensional data, where the number of variables in the dataset is large, and the sample size is relatively small. In such cases standard statistical methods do not perform well, making model parameter estimation potentially problematic. In order to deal with such high dimensional data, statisticians have developed estimators, such as the lasso, that are specially designed to provide model parameter estimates for such scenarios. Recently, this work has been extended to the context of high dimensional multilevel, or mixed effects data in which individuals at level-1 are nested within clusters at level-2. Such data structures are extremely common in the social sciences, particularly education and sociology. The goal of this simulation study was to assess a multilevel extension of the lasso estimator in high dimensional multilevel data case, and to compare this approach with the standard restricted maximum likelihood estimator typically used to fit multilevel models. Results of the study demonstrated that the multilevel lasso yielded better control of the Type I error rate and better parameter coverage than did REML, when level-1 and level-2 sample sizes were small, and there were many predictor variables. Implications of these results are discussed.

In some research and evaluation contexts, the number of variables that can be measured (p) approaches, or even exceeds the number of individuals on whom such measurements can be made (N). For example, researchers working with individuals identified with a rare psychological diagnosis may have difficulty obtaining individuals for their research samples. Once such people are found the researcher may wish to make a relatively large number of cognitive and affective measurements for each participant. The result of small samples coupled with a large number of measurements is commonly referred to as high dimensional data. With such a limited sample size in conjunction with a large number of variables, standard statistical models such as regression, which could normally be used to address the research questions may not work well. Specifically, in the high dimensional context such models tend to yield biased standard errors for the model coefficient estimates (Bühlmann & van de Geer, 2011). A direct consequence of these biased standard errors are inaccurate Type I error and power rates for the tests of the null hypothesis that the coefficient is not 0 in the population. These problems may in turn lead the researcher to erroneous inferences regarding relationships among the independent and dependent variables. In addition, as noted above high dimensionality can also result in parameter estimation bias due to the presence of collinearity, or very strong relationships among the independent variables (Fox, 2016). This model parameter estimation bias can result in potential misrepresentations of the relationships among the variables in the model. Finally, when p in fact exceeds N, it is simply not possible to obtain LS estimates for the model parameters, and the researcher is not able to address the research questions of interest.

The goal of this study is to describe a statistical methodology designed specifically for dealing with high dimensional data in the context of multilevel and mixed effects models. Please note that throughout this manuscript I will use these terms interchangeably to refer to a set of

models involving data structures at multiple levels, as is described in more detail below. These models are becoming increasingly popular in the fields of psychology and education, and as such are being used in a wide variety of applications, some of which involve multilevel data structures. Recently, Schelldorfer, Bühlmann & van de Geer (2011) described an extension of the well known least absolute shrinkage and selection operator (lasso) for use with multilevel data. The purpose of the current study is to investigate the performance of the multilevel lasso through the use of a Monte Carlo simulation. This work extends earlier simulation work by Schelldorfer, et al. (2011), which was fairly limited in scope, and which focused primarily on data scenarios more commonly seen in genetics research than in the social sciences (i.e., 300 to 1000 independent variables, and samples of 150 and 180). The simulation study used here was based upon a review of the multilevel data literature in the social sciences and includes a wider array of conditions with respect to the number of independent variables and the sample sizes. The remainder of the manuscript is organized as follows: First, multilevel models are briefly described in order to provide context for the subsequent discussion of methods. Next, the lasso and multilevel lasso are described, followed by a discussion of the study goals. The methodology used to assess these goals is then described in detail, followed by a presentation of the results. Finally, the implications of these results are discussed, and recommendations for practice are described for researchers.

Multilevel models

Multilevel models (MLMs) are used in the analysis of data in which individuals (level-1) are nested within clusters (level-2), and the clusters could themselves be nested within higher order clusters (level-3). MLMs can also be used in the case of longitudinal data, where measurements taken at different points in time are nested within the individuals on whom they

were made. As mentioned previously, with multilevel data the correct modeling of the relationship between a dependent variable and one or more independent variables must account for the nested structure in order to ensure that estimation bias for parameters and their standard errors is eliminated (Snijders & Bosker, 2012). One of the most common MLMs is the random intercept model, which takes the form:

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij} \quad (1)$$

Where

y_{ij} =Dependent variable value for individual *i* in cluster *j*

β_{0j} =Intercept for cluster *j*

β_1 =Slope relating independent variable *x* to dependent variable *y*

x_{ij} =Value of *x* for individual *i* in cluster *j*

ε_{ij} =Random error for individual *i* in cluster *j*

In turn, β_{0j} can be expressed as

$$\beta_{0j} = \gamma_{00} + U_{0j} \quad (2)$$

Where

γ_{00} =Mean intercept across clusters

U_{0j} =Unique effect of cluster *j* on the intercept

The parameter γ_{00} is referred to as a fixed effect, meaning that it takes the same value for all clusters, and U_{0j} is a random effect that varies across clusters. As an example, for students nested within schools this would mean that β_{0j} would differ across schools, including a common component across schools (γ_{00}), as well as a component unique to the individual school (U_{0j}). Essentially, allowing for these varying intercepts in the model is allowing for schools to have unique means on the dependent variable, even while there is a common mean across all schools.

In model (1), β_1 is treated as a fixed effect indicating that it is constant across clusters. In the school research context this would mean that the relationship between the independent and dependent variables is the same for all schools. It is also possible to fit a random coefficients model in which β_1 has both fixed and random components, just as we have here for β_{0j}, thereby allowing for different relationships between the independent and dependent variables across schools. The error term, ε_{ij}, is a random effect and assumed to be normally and independently distributed across individuals within the same cluster, with $\varepsilon_{ij} \sim N(\mathbf{0}, \Lambda_j)$. Likewise, $U_{0j} \sim N(\mathbf{0}, \Psi)$, and is assumed to be independent across clusters.

The model parameters in (1) and (2) are typically estimated by maximum likelihood (ML) or restricted ML (REML) estimation. With regard to estimating the model parameters themselves (β_1, γ_{00}), ML and REML provide essentially identical results. However, they differ in terms of how the standard errors of these parameters are calculated. Specifically, the degrees of freedom used in ML do not account for the fact that the parameters themselves are being estimated, leading to a negative bias in the standard error estimates (Kreft & de Leeuw, 1998). In contrast, REML standard error estimates do use degrees of freedom that account for the estimation of the model parameters, thereby producing unbiased estimates (Snijders & Bosker, 2012; Lindstrom & Bates, 1988). REML was used in the current study.

The lasso

As noted earlier in the manuscript, high dimensional data can lead to problematic estimation using standard methods, including REML (Schelldorfer, Bühlmann, and van de Geer, 2011). As a result of these issues in the context of standard single level data structures, statisticians have worked to develop estimation methods that can better handle high dimensional data. One such approach is known collectively as regularization or shrinkage methods. These

regularization methods have in common the application of a penalty to the standard least squares estimator that is commonly used to fit a variety of linear and nonlinear regression models. The penalty is devised in such a way that the coefficients linking the independent variables to the dependent variables are made smaller, or shrunken, so that only those that are most strongly related to the dependent variable are retained in the model, whereas the others are eliminated by having their coefficients reduced to 0. The goal of this methodology is to eliminate from the high dimensional model many of the independent variables that exhibit weak relationships to the dependent variable, and thus render the resulting model non-high dimensional; i.e., with only a few salient variables rather than the very large number included in the original model. One of the first such regularization approaches developed for this purpose was the least absolute shrinkage and selection operator (lasso; Tibshirani, 1996). The fitting criterion for the lasso is written as

$$e^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}\left|\hat{\beta}_j\right| \quad (3).$$

Where

y_i =The observed value of the dependent variable for individual *i*

$\hat{y}_i$ =The model predicted value of the dependent variable for individual *i*

$\hat{\beta}_j$ =Sample estimate of the coefficient for independent variable *j*

λ =Shrinkage penalty tuning parameter

The tuning parameter, λ, is used to control the amount of shrinkage (i.e. the degree to which the relationship of the independent variables to the dependent variable are down weighted or removed from the model). Larger λ values correspond to greater shrinkage of the model; i.e. a greater reduction in the number of independent variables that are likely to be included in the final model. On the other hand, a λ of 0 leads to the least squares estimator. Given the goal of

minimizing e^2, the parameter estimates ($\hat{\beta}$) will be reduced in size, and some will even be set to 0, while at the same time the predictions ($\hat{y}$) based upon the parameter estimates should be as accurate as possible, meaning that the parameter estimates cannot all be minimized or set to 0. In other words, the goal of the lasso estimator is to eliminate from the model those independent variables that contribute very little to the explanation of the dependent variable, by setting their $\hat{\beta}$ values to 0, while at the same time retaining independent variables that are important in explaining *y*.

A key aspect of successfully using the lasso is determining of the optimal λ value. A number of approaches for this purpose have been recommended in the literature, including using cross-validation to minimize the mean squared error (Tibshirani, 1996), and selecting the value of λ that minimizes the Bayesian information criterion (BIC). This latter approach was recommended by Schelldorfer, et al. (2011), and was found by them to work well for selecting the optimal tuning parameter value. In addition, work by Zou and Yu (2006) also supported the use of the BIC for this purpose. Therefore, in the current study the BIC was used to select the optimal value of λ. Essentially, a large number of potential λ are tried, the model using each is fit to the data and the BIC values for the models are compared, with the one yielding the smallest BIC being selected.

Multilevel lasso

Schelldorfer, Bühlmann, and van de Geer (2011) described an extension of the lasso estimator that can be applied to multilevel models. This multilevel lasso (MLL) involves the same basic penalty function as described in equation (3), but with additional terms to account for the variance components that are an integral part of the multilevel model in equations (1) and (2). Specifically, the MLL estimators minimizes the following function:

$$Q_\lambda(\beta, \tau^2, \sigma^2) \coloneqq \frac{1}{2} ln|V| + \frac{1}{2}(y_i - \hat{y}_i)'^{V^{-1}(y_i - \hat{y}_i)} + \lambda \sum_{j=1}^{p} \left|\hat{\beta}_j\right| \quad (4)$$

Where

τ^2 =Between cluster variance at level-2

σ^2 =Within cluster variance at level-1

V =Covariance matrix

Thus, the estimates of the model parameters are obtained with respect to the penalization of the level-1 coefficients. In all other respects, this estimator works similarly to the standard lasso of equation (3), including through the use of BIC to select the optimal value of λ.

Standard error estimation for MLL

In order to conduct inference for the MLL model parameters, standard errors must be estimated. Currently, the algorithm does not include standard error estimation. Therefore, in order to allow for the calculation of confidence intervals for each model parameter, an alternative approach must be used. It is proposed here that the block bootstrap methodology might serve as an effective means to calculate standard error estimates for each model parameter, and thereby allowing for the calculating confidence intervals allowing for inference. Traditionally, application of the bootstrap has involved the sampling with replacement of individual members of the sample. However, in the context of multilevel data the block bootstrap approach has been proposed such that, rather than resampling individuals themselves, clusters or blocks of individuals corresponding to their organizing unit (e.g. classrooms or schools) are resampled instead. Block bootstrapping has been used with multilevel data to estimate standard errors from survey data (Kovacevic, Rong, & You, 2006), to correct standard errors in linear regression (Cameron, Gelbach, & Miller, 2008), to calculate standard errors for multilevel DIF assessment

(French, Finch, & Valdivia Vazquez, 2016), and to estimate standard errors for dynamic factor analysis (Zhang & Browne, 2010). The block bootstrap involves the following steps:

1. Calculate the test statistic of interest (e.g., β_j) for the original sample.
2. Resample *m* blocks of individuals with replacement, where *m* is equal to the number of distinct level-2 (e.g. schools) in the sample, where *m* is the number of level-2 units.
3. For each bootstrap sample calculate the parameter estimate; i.e., coefficient.
4. Repeat this procedure *B* (e.g. 1000) times.
5. Calculate the bootstrap standard error as:

$$S_\beta = \sqrt{\frac{\sum_{b=1}^{B}(\beta_0 - \bar{\beta})^2}{B-1}} \quad (5)$$

Where

$\bar{\beta}$=Mean coefficient estimate across the *B* bootstrap samples

The standard error from equation (5) can then be used to construct a confidence interval for β_0 as:

$$\beta_0 \pm 1.96 S_\beta \quad (6).$$

This is the methodology used in the current study.

Study goals and hypotheses

The primary goal of this study was to investigate the performance of the lasso estimator in the context of high dimensional multilevel data. Previous authors (Bühlmann & van der Geer, 2011; Hastie, Tibshirani, & Friedman, 2009; Zou & Hastie, 2005; Tibshirani, 1996) have described how in the single level regression context the lasso is able to control the Type I error rate for tests of the relationships between independent and dependent variables in cases where standard estimators are not. Though this prior work has found that the lasso estimates do exhibit some bias, the level of bias is relatively small (e.g., Tibshirani). Therefore, based on this earlier work with the lasso in the single level data context it is hypothesized that the lasso will control the Type I error rate better than the standard REML estimator for situations involving high

dimensional data. In addition, it is also hypothesized that the lasso estimates will exhibit more negative parameter estimation bias than their REML counterparts, and that power for the lasso will be lower than that of REML, particularly for small sample sizes and with more independent variables. Finally, it is hypothesized that given the more accurate standard error estimates expected for the lasso, coverage rates for this estimator will be closer to the nominal 0.95 level than will be those of the REML estimator.

Method

The aforementioned goals of this study were addressed using a Monte Carlo simulation study with 1000 replications per combination of conditions, which are described below. Data were generated from a 2-level random intercept linear model, as in equation (1), using Mplus, version 7.11 (Muthén & Muthén, 2015). REML estimation was carried out using the R package `lme4` (Bates, Maechler, Bolker, & Walker, 2015), whereas MLL estimates were obtained using the `lmmlasso` function in the `lmmlasso` library (Schelldorfer, 2015). For each replication dataset, the shrinkage parameter was determined based upon the value of the BIC, as described above. Standard errors for the MLL estimates were obtained using the block bootstrap, as was described above. The focus of the simulation was on the level-1 predictors. The data generating conditions that were manipulated in this study are described below.

Level-1 and level-2 sample sizes

The simulated sample sizes per cluster were 5, 10, and 20, and the number of clusters that were simulated were 5, 10, 20, 30, 50, and 100. These values were selected so as to reflect a variety of total sample size conditions, from very small (25 total) to large (2000 total). In addition, these values were selected based upon the results of prior research examining the relationship of sample size and parameter estimation in the context of multilevel modeling. For

example, based on the work of Kreft (1996), Snijders and Bosker (2012), and Hox (2010), it has been suggested that somewhere between 20 and 50 level-2 units should be present when data analysts use the REML estimator. Thus, it was of interest to ascertain the performance of MLL and REML in cases where the number of level-2 units fell below these guidelines, and in cases where the number of level-2 units was well in excess of these values. In addition, the impact of high dimensionality was a primary focus here, and therefore the total sample size on the performance of both estimators was also of interest, and thus a wide array of values were simulated in this study.

Number of level-1 and level-2 independent variables

In addition to the sample size, the number of independent variables at both level-1 and level-2 were also manipulated in this study. In the low dimensionality case, 2 predictors were included at level-1 and 1 predictor level-2 was included at level-2, and in the high dimensionality condition there were 10 level-1 predictors and 5 level-2 predictors. These values were selected in order to reflect a range of conditions that might be expected in the social and behavioral sciences. An examination of 40 studies published in psychology journals in 2017 that used multilevel modeling revealed that the average number of level-1 predictors was 4.4, with a maximum of 8. The average number of level-2 predictors used in these studies was 1.6. Although it is recognized that this is merely a snapshot of the research in the literature, it is believed that these are representative numbers of the level-1 and level-2 predictors, respectively. Thus, the current study was designed to include values at the low and high ends of what is seen in the published psychology literature. The correlations among the independent variables was set equal to 0.3 across conditions, in order to reflect a moderate relationship among them.

Number of coefficients with non-0 population values

The number of independent variables that were simulated to have a relationship with the dependent variable was also manipulated in this study. The purpose for including this condition was to assess the Type I error and power rates for the two estimators under a variety of conditions. For both the low and high dimensional cases, the number of non-0 coefficients was manipulated. In the low dimensional case, one set of simulations was conducted in which all of the independent variables at both levels were simulated to have coefficients of 1 in the population. In the other set of conditions for the low dimensional case, one of the level-1 predictors had a coefficient of 1 in the population, and the other had a coefficient of 0, as did the level-2 predictor. In the high dimensional condition, one set of simulations was such that all of the level-1 and level-2 predictors had coefficients of 1 in the population, whereas for the other set of simulations 5 of the 10 level-1 predictors were simulated to have coefficients of 1, and 5 to have coefficients of 0. In this latter set of conditions, two of the 5 level-2 coefficients were simulated to have coefficients of 1, and the other three to have coefficients of 0.

Intraclass correlation (ICC)

Two values of the ICC were simulated in this study, 0.05 and 0.33. These values were selected because they represent a very small impact of the level-2 units on the outcome (0.05), and a relatively large such impact (0.33).

Outcome variables

There were several outcomes of interest in this study, including parameter estimation bias, the standard error of the estimates, coverage rates for the estimates, Type I error rate, and power, all for the level-1 predictor. Specifically, one of the level-1 predictors was selected as the target, and results are presented below for that target variable. Results for the other level-1 predictors were examined and compared to those for the target, and were found to be extremely

similar to those of the target. Thus, results for the target were the only ones included in the results in order to keep the results at a manageable length. The parameter estimation bias was calculated as:

$$bias = \hat{\theta} - \theta \quad (7)$$

Where

$\hat{\theta}$ =Parameter estimate

θ =Data generating value.

The standard error of the estimates was calculated empirically using the following equation:

$$\sqrt{\frac{\sum_{r=1}^{R}\left(\hat{\theta}-\theta\right)^2}{R-1}} \quad (8)$$

Where

R =Total number of replications; e.g., 1000.

The coverage rate was the proportion of replications for which the 95% confidence interval constructed using the sample data included the data generating value of the parameter. Therefore, if an estimator is working appropriately, the coverage value should be 0.95. The Type I error rate is simply the proportion of replications for which the null hypothesis $H_0: \theta = 0$ was rejected when it should not have been. Likewise, power was the proportion of replications for which this null hypothesis was rejected when it should have been rejected.

In order to identify the main effects and interactions of the manipulated study factors that were related to each outcome, analysis of variance (ANOVA) was used, along with the partial η^2 effect size. For an effect to be considered meaningful in the context of this study, it needed to be both statistically significant, and to have $\eta^2_{Partial}$ value of 0.1 or greater. This latter condition was used because it would mean that the main effect or interaction accounted for at least 10% of the variation in the study outcome.

Results

Parameter estimation bias

With respect to the amount of parameter estimation bias, ANOVA results identified the interaction of estimation method by number of groups by sample size per group as the highest order statistically significant effect $\left(F_{10,13} = 4.453, p = 0.007, \eta^2_{Partial} = 0.774\right)$. All other significant effects were subsumed within this interaction. Estimation bias by number of groups, sample size per group, and estimation method appears in Table 1. Based upon these results, it can be concluded that the degree bias was greater for REML than for the MLL when the number of groups was 5, or there were 10 groups and the sample size per group was 5. For example, at the 5 groups 5 individuals per group condition for REML bias was more than 10 times larger than was the case for MLL. For all other conditions, however, the two methods yielded very comparable, and very low levels of estimation bias. In addition, this pattern of results was present regardless of the population value for the coefficient (0 or 1).

Standard error

ANOVA results showed that the interaction of estimation method by number of groups by sample size per group was the highest order statistically significant such term $\left(F_{10,13} = 2.916, p = 0.037, \eta^2_{Partial} = 0.692\right)$, with all other significant model terms being subsumed in this interaction, or not statistically significant. The standard errors by method, number of groups, and sample size per group appear in Table 2, and reveal that the standard errors for the two approaches are very comparable across most of the simulated conditions. The lone exception to this pattern occurred when data were simulated for 5 groups, with a sample size of 5 individuals per group, in which case REML had a larger standard error than did that

produced by MLL. Otherwise, standard errors for the two methods were within 0.001 of one another across conditions.

Coverage

As with the bias and standard error outcomes, ANOVA was used to ascertain which of the manipulated factors in the simulation study were associated with the coverage rates for the model parameters. It was found that the interaction of number of groups, sample size per group, and estimation method were associated with coverage $\left(F_{10,13} = 2.708, p = 0.050, \eta^2_{Partial} = 0.673\right)$. Coverage rates by method, number of groups, and number of items appear in the bar chart in Figure 1. A reference line has been placed at the 0.95 value on the graphs, denoting the nominal coverage level. Thus, when a method is working appropriately with respect to coverage, the bar should meet this line. In fact, for REML the coverage rates were consistently below the nominal level when there were 5 or 10 groups, regardless of the sample size per group. In addition, for 20 and 30 groups, the REML coverage rates were below 0.95 for samples of 5 individuals per group. In contrast, the coverage rates for the MLL approach were always at or slightly above the nominal 0.95 level.

Type I error rate

When the level-1 coefficients for variables were simulated to be 0 (i.e., there was no relationship between the independent variable and the response), a statistically significant result would represent a Type I error. In order to determine which of the manipulated factors were associated with the Type I error rate, ANOVA was used, as mentioned in the methods section. The interaction of the number of groups by the estimation method was found to be statistically significantly related to the Type I error rate $\left(F_{5,8} = 10.562, p = 0.002, \eta^2_{Partial} = 0.868\right)$. Figure 2 displays the Type I error rate for each estimation method by the number of groups.

Note that there is a reference line at the nominal 0.05 Type I error level. In addition, per recommendations from Bradley (1978) error rates between 0.025 and 0.075 were considered to be in control. Perhaps the most obvious result made apparent in Figure 2 is that the REML Type I error rate was out of control when the number of groups was 20 or fewer, whereas for 30 or more groups the error rate was in control. In contrast, the Type I error rate for MLL was always in control, and indeed well below 0.05, for all number of groups conditions.

Power

ANOVA results for the power for detecting model parameters that are not 0 in the population identified the interaction of number of level-2 groups by the sample size per group by estimation method $\left(F_{10,13} = 15.853, p < 0.001, \eta^2_{Partial} = 0.924\right)$ as being statistically significantly related to the power rate. The power by number of level-2 groups, sample size per group, and estimation method appear in Figure 3. From these results, it can be seen that power for REML was lower than that for MLL with 20 or fewer groups, regardless of the number of individuals in each group. In addition, this effect of the number of groups was magnified by the sample size per group, so that when groups were smaller, the power differential between the two methods was greater. Under this combination of conditions, power for MLL was always between 0.98 and 1. For 30 or more level-2 groups, power for the two methods was very comparable and always between 0.98 and 1.

Discussion

Researchers in some areas of the social sciences will sometimes face the situation in which they have relatively small samples and a relatively large number of variables of interest. In such cases, standard parameter estimation algorithms, such as ordinary least squares and maximum likelihood will not provide stable or reliable estimates. In an attempt to address this

problem, Tibshirani (1996) introduced the lasso estimator, which was designed to reduce the effective set of predictors in a model to include only those which are most strongly associated with the dependent variable of interest. In the context of least squares estimation the lasso has been shown to produce shrunken estimates, which tend to be somewhat negatively biased, though the degree of bias is typically small. In addition, the lasso, and other shrinkage estimators, have been found to control the Type I error rate better than do standard estimators such as least squares and maximum likelihood (Zou & Hastie, 2005). In short, for single level data shrinkage methods such as the lasso have been found to work well in terms of yielding reasonably accurate parameter estimates while also controlling the Type I error rate.

High dimensionality is not a problem limited to single level models, and for that reason the current study was designed to explore the performance of the MLL for use with multilevel data structures. Results of the study generally showed that in the context of a 2-level random intercept model, the MLL estimator is a very viable alternative to the standard REML approach most commonly used in practice, even when the data were not high dimensional. Specifically, when the sample sizes were small at both levels 1 and 2, MLL yielded less biased parameter estimates than did REML, and for larger samples both methods had very similar (and extremely small) levels of estimation bias. Similarly, for the smallest sample size conditions, MLL yielded more controlled Type I error rates, and higher power than did REML. The results of this study also demonstrated that use of the block bootstrap for estimating the standard errors of the level-1 parameter estimates is also viable, as these standard errors were generally very similar to those of REML for larger sample sizes, and somewhat smaller when the sample sizes were small. This use of the block bootstrap represents an extension to the work of Scheldorfer, et al. (2011), Finally, the coverage rates for the MLL estimator using the block bootstrap to estimate the

standard error were always at or slightly above the nominal 0.95 level, whereas REML had coverage rates below the nominal level whenever the number of level-2 clusters was less than 20, and these were lowest for the combination of five level-1 and five level-2 units.

With regard to the research hypotheses described above, several conclusions can be reached. First, as was hypothesized based on earlier work, MLL was better able to control the Type I error rate than was REML. In addition, the hypothesis that the MLL coverage rates would be better than those of REML was also supported by this study. However, the hypotheses that power would be lower for MLL was not supported, nor was the hypothesis that its parameter estimates would be more negative than those of REML. These latter results are likely at least partially a result of the fact that the MLL loss function in equation (4) involves not only the fixed effects (i.e., β_1) but also the two variance component terms. Thus, when determining the penalty for the coefficients, MLL accounts not only for the level-1 predictors but also for the variation both within and between level-2 clusters. For the standard single level lasso, the loss function is only influenced by the degree of disparity between the observed and model predicted dependent variable values. However, for MLL not only are these terms important, but so are the variance components estimates. This fact would appear to largely mitigate the impact of the shrinkage process on the estimates. Interestingly, the numbers of level-1 and level-2 predictor variables were not found to be related to the performance of the estimators. Thus, in the context of multilevel data, it would appear that the sample sizes at levels 1 and 2 may be more salient in terms of estimation performance than are the number of independent variables.

Directions for future research

The results of this study appear to support the performance of the MLL estimator with small samples and high dimensional data. In addition, they buttress earlier work by Scheldorfer,

et al. (2011) suggesting that this estimator may be particularly useful with relatively small samples. Despite these results, however, there does remain room for further research. This study was intended to represent the first fairly large scale simulation examination of the performance of MLL. However, more work needs to be done in this regard. First, a wider array of number of level-1 and level-2 predictors needs to be examined. The values selected for this study were taken from published literature in the social sciences. However, more extreme numbers of predictors should be examined in future work, perhaps with as many as 20 to 30 such variables. Work by Scheldorfer, et al., examined very extreme dimensionality with 300 to 1000 independent variables, and found that MLL worked well in terms of parameter recovery. When coupled with the results presented in the current study, these earlier results are certainly suggestive that MLL should work well with 20 or 30 predictors. However, by themselves these very large numbers are not particularly informative for most social science research, as the number of predictors will typically not be in the several hundreds. In addition to including more independent variables, future work should also examine a wider range of ICC values, and a wider range of sample size conditions. In particular ICC values in between the small (0.05) and large (0.33) values included here could be informative for applied researchers. Finally, future research should examine the performance of MLL and REML in terms of estimating variance components estimation in the high dimensional case. In order to keep the current study well focused, only the level-1 predictor estimates were examined. This decision was made because these estimates are typically of primary interest to researchers using MLLs. It is hoped that this focus allowed for a clear picture to be developed regarding the performance of the two estimators in the high dimensional case. Now that such work has been completed, a next step would be to investigate the estimation of the variance components themselves.

Conclusions

As noted earlier, the current study was designed to build on the work by Scheldorfer, et al. (2011) with regard to the performance of the MLL estimator in the context of high dimensional data. The results of this study have found that MLL is indeed a viable alternative to REML, not only in the context of high dimensional data but indeed for general use. MLL always performed as well as REML in the simulated conditions, and was preferable for small sample sizes. Thus, researchers are encouraged to consider using it whenever the level-2 sample size is less than 30. MLL is particularly useful in terms of controlling the Type I error rate of the level-1 predictors. In addition, MLL appears to perform as well as REML for larger sample sizes, so that it would be appropriate to use even in cases where the sample sizes at levels 1 and 2 are relatively large. Finally, the results of this study show that the block bootstrap is an appropriate method for estimating the standard error of the level-1 estimates. This was not an issue described by Scheldorfer, et al., and one which adds to the current use of this estimator.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software, 67(1),* 1-48.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin: Springer-Verlag.

Cameron, A.C., Gelbach, J.B., & Miller, D.L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics, 90(3),* 414-427.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models.* Thousand Oaks, CA: Sage.

French, B. F., Finch, W. H., & Valdivia Vazquez, J. A. (2016). Predicting Differential Item functioning on Mathematics Items using multilevel SIBTEST. *Psychological Test and Assessment Modeling, 58,* 471-483.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer-Verlag.

Hox, J.J. (2010). *Multilevel Analysis: Techniques and Applications.* New York: Rougledge.

Kovacevic, M.S., Rong, H., & Yon, Y. (2006). Bootstrapping for Variance Estimation in Multi-Level Models Fitted to Survey Data. Paper presented at the annual meeting of the American Statistical Association, Seattle, WA.

Kreft, I. (1996). *Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies.* California State University, Los Angeles.

Kreft, I. & de Leeuw, J. (1998). *Introducing Multilevel Modeling.* London: Sage.

Lindstrom, M.J. & Bates, D.M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical*

Association, 83(404), 1014-1022.

Muthén, L. K., & Muthén, B. O. (2015). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

Schelldorfer, J. (2015). R Package: Linear Mixed-Effects Models with Lasso.

Schelldorfer, J,, Bühlmann, van de Geer, S. (2011). Estimation for High-Dimensional Linear Mixed-Effects Models using l1-Penalization. *Scandinavian Journal of Statistics, 38,* 197-214.

Snijders, T.A.B. & Bosker, R.J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Los Angeles: Sage.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B., 58,* 267-288.

Zhang, G. & Browne, M.W. (2010). Bootstrap Standard Error Estimates in Dynamic Factor Analysis. *Multivariate Behavioral Research, 45,* 453-482.

Zou, H. & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B., 67(2),* 301-320.

Table 1: Parameter Estimation Bias by Estimation Method, Number of Level-2 Groups, and Sample Size per Group

Groups	Sample Size per Group	REML	MLL
5	5	.1020	.0185
	10	.0608	.0190
	20	.0457	.0145
10	5	-.0277	-.0181
	10	-.0090	-.0092
	20	.0007	.0009
20	5	.0034	.0034
	10	.0014	.0014
	20	.0028	.0027
30	5	-.0031	-.0029
	10	.0063	.0061
	20	.0003	.0004
50	5	-.0009	-.0009
	10	-.0075	-.0074
	20	-.0008	-.0009
100	5	-.0062	-.0062
	10	.0005	.0006
	20	-.0004	-.0004

Table 2: Parameter Estimate Standard Error by Estimation Method, Number of Level-2 Groups, and Sample Size per Group

Groups	Sample Size per Group	REML	MLL
5	5	.3775	.2813
	10	.2135	.2135
	20	.1318	.1325
10	5	.2181	.2187
	10	.1352	.1355
	20	.0841	.0840
20	5	.1418	.1418
	10	.0909	.0909
	20	.0588	.0588
30	5	.1162	.1163
	10	.0736	.0734
	20	.0540	.0538
50	5	.0812	.0809
	10	.0560	.0561
	20	.0423	.0425
100	5	.0599	.0600
	10	.0420	.0421
	20	.0298	.0339

Figure 1: Coverage Rates by Estimation Method, Number of Level-2 Groups, and Sample Size per Group

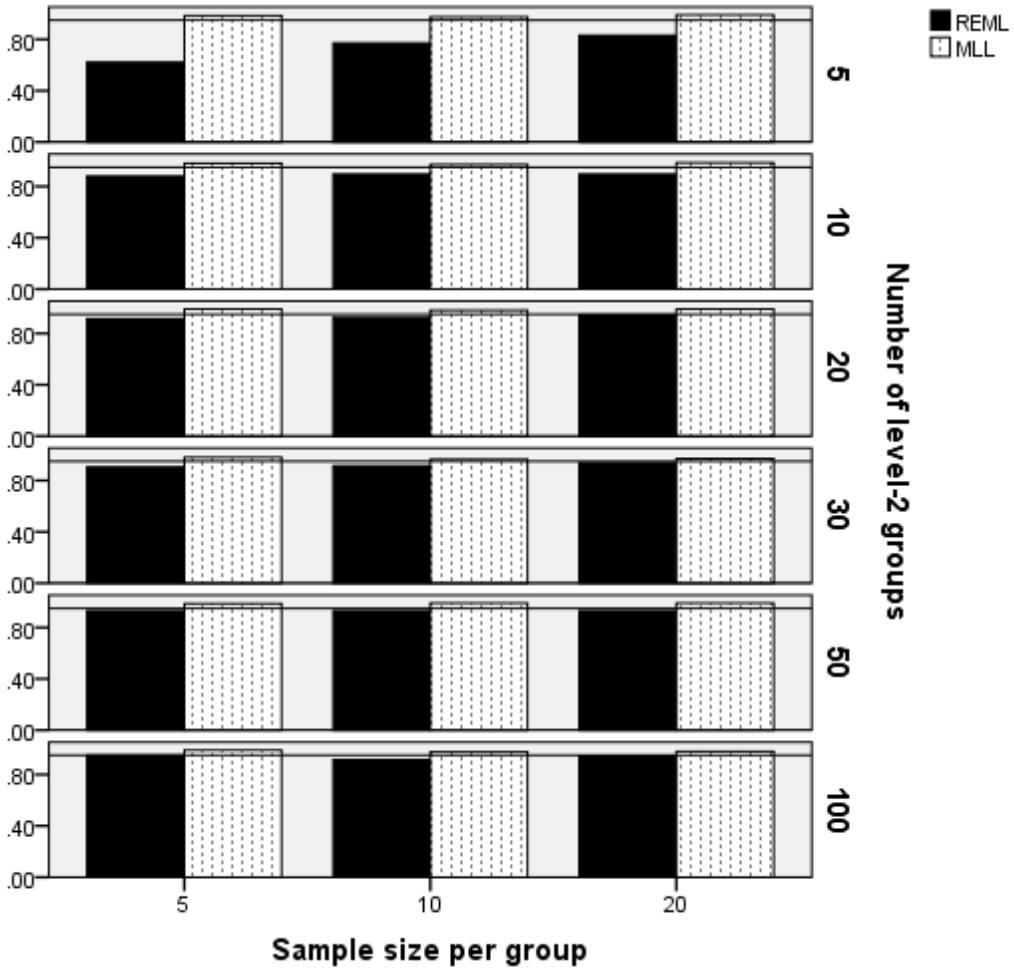

Figure 2: Type I Error Rate for Parameter Estimate by Estimation Method and Number of Level-2 Groups

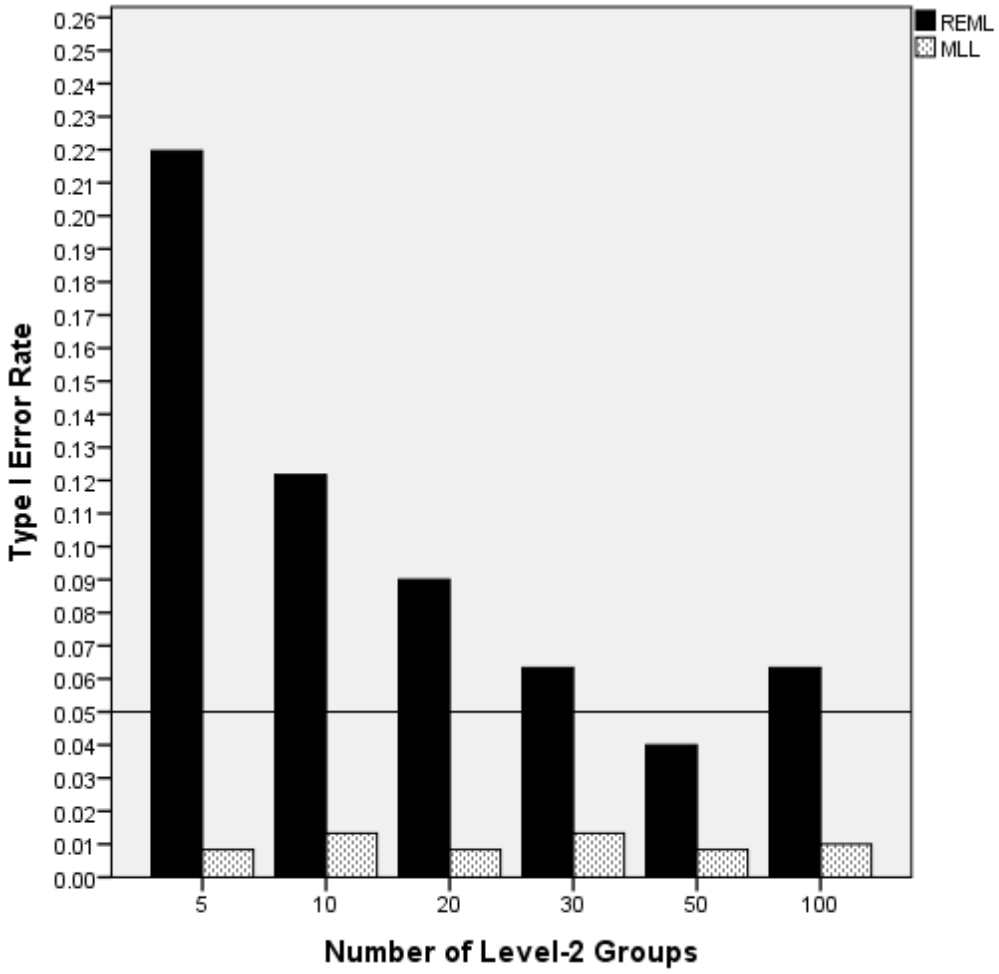

Figure 3: Power for Parameter Estimate by Estimation Method and Number of Level-2 Groups