Taylor & Francis Online
Journals Search Publish
Access provided by The UC Davis Libraries
Login | Register
Home ▶ All Journals ▶ Mathematics, Statistics & Data Science ▶ Journal of the American Statistical Association ▶ List of Issues ▶ Volume 119, Issue 547 ▶ Classified Mixed Model Projections
Enter keywords, authors, DOI, etc. This Journal
Advanced search

Journal of the American Statistical Association ›
Volume 119, 2024 - Issue 547
Submit an article Journal homepage

713
Views
0
CrossRef citations to date
1
Altmetric

Listen
Full access

Theory and Methods

Classified Mixed Model Projections

J. Sunil Rao, Mengying Li & Jiming Jiang

Pages 1805-1819 | Received 01 Jun 2021, Accepted 18 May 2023, Accepted author version posted online: 26 May 2023, Published online: 06 Jul 2023
Cite this article https://doi.org/10.1080/01621459.2023.2218041 Check for updates

Full Article Figures & data References Supplemental Citations Metrics Reprints & Permissions
View PDF View EPUB Share
Formulae display: MathJax

Related research
People also read | Recommended articles | Cited by
Classified Mixed Model Prediction ›
Jiming Jiang et al.
Journal of the American Statistical Association
Published online: 26 Sep 2017
Asymptotic Distribution-Free Independence Test for High-Dimension Data ›
Zhanrui Cai et al.
Journal of the American Statistical Association
Published online: 21 Dec 2023
An Automated Approach to Causal Inference in Discrete Settings ›
Guilherme Duarte et al.
Journal of the American Statistical Association
Published online: 21 Aug 2023
View more

Abstract

In many practical problems, there is interest in the estimation of mixed effect projections for new data that are outside the range of the training data. Examples include predicting extreme small area means for rare populations or making treatment decisions for patients who do not fit typical risk profiles. Standard methods have long been known to struggle with such problems since the training data may not provide enough information about potential model changes for these new data values (extrapolation bias). We propose a new framework called Prediction Using Random-effect Extrapolation (PURE) which involves constructing a generalized independent variable hull (gIVH) to isolate a minority training set which is "close" to the prediction space, followed by a regrouping of the minority data according to the response variable which results in a new (but misspecified) random effect distribution. This misspecification yields "extrapolated random effects" which prove vital to capture information that is needed for accurate model projections. Projections are then made using classified mixed model prediction (CMMP) (?) with the regrouped minority data. Comprehensive simulation studies and analysis of data from the National Longitudinal Mortality Study (NLMS) demonstrate superior predictive performance in these very challenging paradigms. An asymptotic analysis reveals why PURE results in more accurate projections. Supplementary materials for this article are available online.

KEYWORDS:
Extrapolation Generalized independent variable hull Linear mixed model Minority data Regrouping

‹ Previous article View issue table of contents Next article ›

1 Introduction

In many important practical problems, there is interest in the prediction for new data that are outside the range of the training data—the so-called extrapolation prediction or projection problem which standard prediction methods have a difficult time with. Extrapolation as a topic has intrigued researchers back to the time of Archimedes who developed among the first methods for function extrapolation for numerical analysis (Walz 1991). In statistical circles, the focus has been on *predictive modeling* with interesting distinctions being made between predictability and prediction (Ehrenberg and Bound 1993). As noted there, the most interesting prediction or extrapolation problem asks the question if a result found from one set of data will hold for some other, different data (e.g., for a sample from another population). The main issue then boils down to the choice of model form relating the response of interest to a set of available covariates. Typically, even if the training model is correct, extrapolated prediction intervals widen dramatically. If the training model is incorrect, then bias enters the equation as well.

Here is a typical example. An older individual with later stage cancer is deciding whether or not to undergo chemotherapy treatment. His doctors are encouraging him to do so indicating a significantly increased remaining life span over doing nothing. However, the patient also suffers from co-morbidities including significant kidney disease resulting in markedly reduced kidney function and reduced cardiac function that has required him to have had significant bypass surgery. The doctors indicate that their estimate for increased length of remaining life is really not based on a model which could easily account for the co-morbidities and on top of that, is not really designed for patients of advanced ages. So any prediction would represent an extrapolation from the training model. The best advice they can give is to use their estimate as an upper bound. When weighing such big decisions, this proves less than satisfactory. Yet, as we know, this type of problem occurs regularly in practice and frankly reflects a lack of diversity (age, gender, race, etc) in research training data models. If the promise of precision medicine is to be more fully realized, then not only should research involve more diverse populations, but more accurate methods for extrapolation should be developed as well.

This above example is a projection of a response associated with new data and the underlying model could be a linear mean residual life model (Chen and Cheng 2006). In Linear mixed model settings (which is our focus here), the projection of a mixed effect associated with new data is also of interest. For instance in small area estimation (SAE) (Rao and Molina 2015), we use mixed models to derive model-based estimators of area mixed effects, such as small area means, using mixed model prediction (MMP) (Jiang and Lahiri 2006) or classified mixed model prediction (CMMP) (Jiang et al. 2018). When new areas present as outside of the range of the auxiliary variables in the training data, the goal is to generate a mixed model projection of the mixed effect for those areas. In the SAE literature, to date we can point to work by Saei and Chambers (2005b) who looked at out-of-sample estimation for the area level model where they relaxed the independent area effects assumption and allowed random effects to be spatially correlated. The same authors developed an earlier similar approach for the case where unit level data are available from the in-sample areas (Saei and Chambers 2005a).

While older literature has come to the conclusion that extrapolation methods have had mixed success at best Armstrong (1984), as we know, an explosion of work has been done in the area of machine learning or predictive modeling in the last 25 years. These models (which also include ensembles) usually make use of more local structure in the data and are hence more flexible. They are also inherently more computationally intensive than traditional statistical models. However, it has been noted that these models do not remedy the challenges of extrapolation prediction (http://freerangestats.info/blog/2016/12/10/extrapolation).

Nevertheless, there are some ways to make in-roads here. If the training data are grouped in a nested error structure, then we could "map" the test data to one of the more representative groups and use the recently developed classified mixed model prediction (CMMP) approach of Jiang et al. (2018) (more details to follow). Similarly, if we could identify a subset of the training data that are "closer" to the test data, then we can build a training model using these as advocated by Fine (1970). Both of these methods can provide improved mean squared prediction error (MSPE) in some cases. In this study, we will do something quite different. We will identify a set of *minority* data points which are close to the test data and then fit a *regrouped* nested error mixed effects model to these. As we will show, the regrouped random effects are in a sense extrapolated from the full training data random effects. Predictions made using this regrouped structure yield a method we call Prediction Using Random-Effects Extrapolation (PURE).

1.1 Illustrative Example

A simple example illustrates how this method works. It is assumed that, for $i = 1, \ldots, m$, $\mathbf{y}_i^{(k)}$ follow a mixed model as follows:

$$\mathbf{y}_i^{(k)} = X_i^{(k)}\boldsymbol{\beta}_k + Z_i^{(k)}\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad (1)$$

where $\mathbf{y}_i^{(k)} = (y_{ij})_{1 \le j \le n_i^{(k)}}$, $X_i^{(k)} = (x_{ij}^{(k)})_{1 \le j \le n_i^{(k)}}$ is a matrix of known covariates, $Z_i^{(k)}$ is a matrix of known covariates, $\boldsymbol{\beta}_k$ is a p-vector of unknown regression coefficients (the fixed effects), $\mathbf{b}_i$ is q-vector of group-specific random effects, and $\boldsymbol{\varepsilon}_i$ is an vector of errors. The subscript (k) denotes the population k, and $1 \le k \le K$. It is assumed that $\mathbf{b}_i \sim N(0, G)$, $\boldsymbol{\varepsilon}_i \sim N(0, R_i)$ and they are independent, and the covariance matrices G and R_i depend on a vector $\boldsymbol{\psi}$ of variance components. Note $\boldsymbol{\beta}_k$ is different for different population k, and the random effects $\mathbf{b}_i$ are the same across k populations. The total number of observations in each population is $n^{(k)} = \sum_{i=1}^m n_i^{(k)}$, and the overall total population $n = \sum_{i=1}^K n^{(k)}$. Note that $n = \sum_{i=1}^m n_i$ where n_i is the number of observations in the group i. If the data follows (1), people usually fit a one component mixed model that assumed only one set of fixed effects parameters when the true model information is unknown, which results in a convenient but "misspecified" model fit.

Assume new test observations, which follow:

$$y_{n,j} = x_n'\boldsymbol{\beta}_n + z_n'\mathbf{b}_I + \varepsilon_{n,j}, \quad 1 \le j \le n_{\text{new}}, \quad (2)$$

where x_n and z_n are known vectors, and I belongs to one of the m groups. The new errors $\varepsilon_{n,j}$ are independent with mean zero, and variance R_{new} and are assumed independent of the training data. Notice $\boldsymbol{\beta}_n \ne \boldsymbol{\beta}_k, 1 \le k \le K$. The mixed effect we wish to predict is $\theta_n = E(y_{n,j} \mid \mathbf{b}_I) = x_n'\boldsymbol{\beta} + z_n'\mathbf{b}_I$ where $I \in [1, \ldots, m]$ but we do not know which group I belongs to.

Let $K = 2$; then we have $\boldsymbol{\beta}_n \ne \boldsymbol{\beta}_1 \ne \boldsymbol{\beta}_2$, where $\boldsymbol{\beta}_1$ denotes the fixed effects parameter for the *majority* data, and $\boldsymbol{\beta}_2$ denotes the fixed effects parameter for the *minority* data. Assume the training data is composed of two sub-populations: 60% of the training data has $\boldsymbol{\beta}_1 = (0.5, 0.4, 0.3, 0.2, 0.1, 0)'$, and fixed covariates X were generated from a normal distribution with $\boldsymbol{\mu}_1 = (30, 150, 105, 0, 0)'$ and $\Sigma_1 = \text{diag}[40, 40, 40, 1, 1]$; the other 40% of the training data has $\boldsymbol{\beta}_2 = (0.55, 0.45, 0.35, 0.2, 0.1, 0)'$, and fixed covariates X were generated from a normal distribution with $\boldsymbol{\mu}_2 = (90, 250, 140, 0, 0)'$ and the same Σ_1. We will consider $m = 30$ and $n_i = 10, 1 \le i \le m$. For the random effects, we assume a random intercept model with random effect variance $G = 20$. We further assume independent sampling errors with $R = I$.

The first three covariates could for instance represent age, weight, and blood pressure and differ in their effect on the response between the two sub-populations. Note that the minority data are represented in each of the m groups. In real applications, the groups could refer to different hospitals, and in each hospital, there are some elderly patients who are overweight with high blood pressure (that represent minority data).

For the test data, we assumed the matched case where the test data is assumed to come from one of the m groups. For the test data, we generated 100 sets of new observations where each set contains five observations. The test data covariates are simulated from a multivariate normal with $\boldsymbol{\mu}_n = (100, 280, 150, 0, 0)'$ and the variance is the same as Σ_1. Notice that the mean vector is even larger than the minority data. In addition, the test data is assumed to have larger fixed effects parameters, that is $\boldsymbol{\beta}_n = (0.65, 0.20, 0.35, 0.2, 0.1, 0)'$. What this means is that we generated new test data fixed effects and attached training data random effects to them at random thus preserving the grouping structure. Once again, we assume independent sampling errors with $R = I$.

We fit the following models: (i) usual regression prediction (RP) which uses the population average value of the random effects (i.e., zero); (ii) classified mixed model prediction (CMMP) which attaches to a test observation, a classified random effect from the training data associated with a group with minimal prediction error; (iii) a usual linear mixed model using mixed model prediction for test observations (Lmer); (iv) a regrouped linear mixed model in which the training observations are reclustered according to their response values and (v) CMMP with regrouped data. We also then fit the same models using a subset of the training data identified to be closest to the test data based upon only their covariate values (details to follow). This is the data-based identification of minority data. We term CMMP on regrouped minority data as prediction using random effect extrapolation (PURE).We also included random forests and gradient boosted regression trees as competitors. All results are based on averages over 100 simulation runs.

Prediction performance results over the 100 simulation runs are presented in Table 1. The PURE prediction method clearly produces the most accurate predictions of all methods examined. Interestingly, flexible machine learning methods like random forests and gradient boosting are not competitive. This seems to confirm previously found limitations of these methods for extrapolation problems.

Table 1 Summary of MSPE simulation results for prediction of mixed effect for toy example.

Download CSV Display Table

A scatterplot of the data is shown in Figure 1 with the majority, minority, and test (extreme) data points colored differently. To better illustrate how the test data points are outside the range of the training data, a scatterplot with respect to the first principal component is also shown. Clearly, the test data does not overlap with any of the training data. Fitted lines (from one realization of the simulation) running through the test set predicted means using the full data (yellow), the minority data (green) are overlaid on the plot. Clearly, the full data or minority data do not provide an adequate fit when trying to extrapolate to the test data points. Additionally, PURE predictions are shown in blue with a fitted line through them.

Fig. 1 Scatterplot (raw on the left, with respect to the first principal component on the right), with overlaid fitted lines (fitted to predicted test means) from introductory example.

Display full size

Table 2 shows the estimated G and R values averaged over the 100 runs of the simulation with empirical standard errors in parentheses in both the original grouped and regrouped situations. Notice how the regrouping has markedly increased the G/R ratio from the original grouping scenario—in a sense, extrapolating the random effects.

Table 2 Estimated G and R and (SEs in parentheses) for toy example over 100 runs.

Display Table

Clearly PURE prediction is intriguing at the very least. The gains in prediction accuracy for extrapolated projections are enormous in this example. By simply focusing on minority data and regrouping the observations, we see remarkable decreases in prediction error. So in this article we will detail what PURE prediction is and why exactly it works using a detailed asymptotic analysis. The article is organized as follows. We first provide a background of CMMP prediction. We then show how to identify a minority subset using a generalized independent variable hull (gIVH) adapted to the mixed model setting. We then put everything together as PURE prediction and conduct extensive empirical studies using simulations. The next section details an asymptotic analysis of the whole procedure which shows why MSPE can be lowered using PURE. A detailed real data analysis is conducted where the focus of interest is predicting age to start smoking for different racial groups from the National Longitudinal Mortality Study (NLMS). We finally end with some conclusions.

Remark 1.

It is natural to consider using $\bar{y}_n$ as a naive estimator of θ. When n_{new} is large, we certainly expect for the PURE to be close to $\bar{y}_n$. This is a property that resembles what is sometimes called design consistency (e.g., in small area estimation problems)—that is, the model-based predictor is close to the design-based one (see e.g., Jiang and Lahiri 2006). Clearly, for smaller sample sizes, the CMMP should be preferable as it borrows strength from other sources through the model. In addition, covariate information is generated from sometimes expensive studies like large scale surveys and not making use of this information can be difficult to justify.

2 Classified Mixed Model Prediction

In the work of Jiang et al. (2018), they proposed classified mixed model prediction (CMMP) as a means to predict subject level values in a mixed model context. They showed that CMMP can have significantly reduced the MSPE by identifying a group that a new test subject belongs to and then using that classification by attaching to the test observation the corresponding group random effect(s). Consider the linear mixed regression model,

Suppose we have a set of training data $y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i$, and the linear mixed model for the training data is

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad (3)$$

where $\mathbf{y}_i = (y_{ij})_{1 \le j \le n_i}$, $X_i = (x_{ij}')_{1 \le j \le n_i}$ is a matrix of known covariates, Z_i is a design matrix for the random effects, $\boldsymbol{\beta}$ is a p-vector of unknown regression coefficients (the fixed effects), $\mathbf{b}_i$ is q-vector of group-specific random effects, and $\boldsymbol{\varepsilon}_i$ is an vector of errors. It is assumed that $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$ are the same as in (1). The total number of observations $n = \sum_{i=1}^m n_i$.

The particular version of CMMP of interest here attempts to make a classified prediction for the conditional mean value associated with a set of new observations, $y_{n,j}, 1 \le j \le n_{\text{new}}$, where the subscript n refers to "new." For the new observations:

$$y_{n,j} = x_n'\boldsymbol{\beta} + z_n'\mathbf{b}_I + \varepsilon_{n,j}, \quad 1 \le j \le n_{\text{new}}, \quad (4)$$

where x_n and z_n are known vectors, and I either belongs to one of the m groups, or I does not match to any of the m groups. The new errors $\varepsilon_{n,j}$ are independent with mean zero, and variance R_{new}. Notice that here $\boldsymbol{\beta}$ is assumed to be the same for the test observations.

We wish to predict the mixed effect associated with the new observations at x_n:

$$\theta_n = E(y_{n,j} \mid \mathbf{b}_I) = x_n'\boldsymbol{\beta} + z_n'\mathbf{b}_I. \quad (5)$$

We can construct the best predictor under the linear mixed model for each group as in Jiang et al. (2018). Estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\psi}}$ of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ can be obtained through Maximum Likelihood (ML), Restricted Maximum Likelihood (REML), or other available approaches like the Best Predictive Estimator (BPE) (Jiang, Nguyen, and Rao 2011). By (5), it follows that the estimator of mixed effects for each group is $\tilde{\theta}_{(i)} = x_n'\hat{\boldsymbol{\beta}} + z_n'\tilde{\mathbf{b}}_i$. Note $\tilde{\mathbf{b}}_i$ is the estimator of the random effects for the group i. For example, in the nested-error regression model (Battese, Harter, and Fuller 1988), $\tilde{\theta}_{(i)} = x_n'\hat{\boldsymbol{\beta}} + \frac{n_i\hat{G}}{\hat{R} + n_i\hat{G}}(\bar{y}_{i\cdot} - \bar{x}_{i\cdot}'\hat{\boldsymbol{\beta}})$ where $\bar{y}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$, $\bar{x}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} x_{ij}$, and $\hat{G}$ and $\hat{R}$ are consistent estimators of G and R.

Then I is classified as i for $i = 1, \ldots, m$ by minimizing the mean squared prediction error (MSPE) of the best predictor for each group defined as

$$\text{MSPE}_i = E\{\tilde{\theta}_{(i)} - \theta\}^2 = E\left\{\tilde{\theta}_{(i)}^2 - 2\tilde{\theta}_{(i)}\bar{y}_n + \theta^2\right\}.$$

The new observations are thus classified into one of the groups i that has minimum observed MSPE_i, which means

$$\hat{I} = \underset{i}{\text{argmin}}\left\{\tilde{\theta}_{(i)}^2 - 2\tilde{\theta}_{(i)}\bar{y}_n\right\}. \quad (6)$$

Then the classified mixed effect predictor is $\hat{\theta}_n = \tilde{\theta}_{(\hat{I})}$. An unmatched version of CMMP was also proposed where the new observations were not assumed to belong to one of the original m groups.

Remark 2.

Jiang et al. (2018) developed a version of CMMP which can be used to predict the response associated with future observations. It assumes the existence of an additional intermediate set of data where both responses and covariates are observed and which area assumed to come from the same group as the new observations for which predictions are desired. More recently Sun et al. (2018) developed a matching procedure which does not require an intermediate dataset. While our focus is on the projection of a mixed effect associated with new observations, the projection of future observations can be handled using the same regrouped mixed model idea (to follow) and then appealing to either Jiang et al. (2018) or Sun et al. (2018) depending on whether intermediate data was available or not.

3 Generalized Independent Variable Hull

Extrapolation is different from interpolation which is a reciprocal problem making predictions base on new data points within the range of observed data point. Thus, the key point to separate extrapolation and interpolation is to define the range of observed data. Conn, Johnson, and Boveng (2015) proposed one possible definition of "the range of observation data" which turns to early works on outlier detection in simple linear regression analysis. Cook (1979) referred that the smallest convex set containing all design points of a full-rank linear regression model as the independent variable hull (IVH). A linear regression model can be written as

$$y = X\beta + \varepsilon, \quad (7)$$

where y is the n-vector observed response, X is the $n \times p$ design matrix, and ε is an n-vector of independent and identically distributed Gaussian errors. Let H be the hat matrix, $H = X(X^TX)^{-1}X^T$, and h denote the maximum diagonal element of H. Then, a new design point, x_0 is within the IVH whenever

$$x_0(X^TX)^{-1}x_0^T \le h. \quad (8)$$

The IVH definition is based on linear model (7) which require full rank of design matrix and iid Gaussian error. Therefore, it can not be applied to generalized models such as binary response or random effects. Cook (1979) notes that design points with maximum prediction variance will be located on the boundary of IVH, then Conn, Johnson, and Boveng (2015) defined a generalized independent variable hull (gIVH) as a set of all predicted locations $\mathbf{S}_0$ for which

$$\text{var}(\lambda_i) \le \max(\text{var}(\lambda_\mathbf{S})), \quad (9)$$

where $i \in \mathbf{S}_0$, λ_i corresponds to the mean prediction at i, $\mathbf{S}$ denoted the set of locations where data are observed, and $\lambda_\mathbf{S}$ denotes predictions at $\mathbf{S}$. Conn, Johnson, and Boveng (2015) proposed that the gIVH can be applied to determine whether predictions are interpolations (predictive design points lying inside the gIVH) or extrapolations (predictive design points lying outside the gIVH). This uses the generalization,

$$\mu = X_{\text{aug}}\beta_{\text{aug}}, \quad (10)$$

where X_{aug} is an augmented design matrix to accommodate the random effects design matrix Z and β_{aug} is the corresponding regression parameter vector. We can then write the prediction variance as,

$$\text{var}(\hat{\lambda}) = \text{var}(\hat{\mu}) = X_{\text{aug}}\text{var}(\hat{\beta}_{\text{aug}})X_{\text{aug}}'. \quad (11)$$

One possibility is to use a flexible generalized additive model (GAM) (Hastie and Tibshirani 1990) and then estimate the appropriate form of $\text{var}(\hat{\beta}_{\text{aug}})$. If y is not on the linear predictor scale (e.g., generalized linear models outside of the normal model), then the delta method can be used to estimate $\text{var}(\hat{\lambda})$ (Conn, Johnson, and Boveng 2015). Outside of these situations, simulation based methods like bootstrapping can be used to estimate the variance.

Let's illustrate gIVH with one example. Suppose we have training data that follows a linear relationship, where $\beta_0 = (0.5, 0.4, 0.3, 0.2, 0.1)'$, and design matrix X were generated from a normal distribution with $\mu_0 = (30, 150, 105, 0, 0)'$, $\Sigma_0 = \text{diag}[40, 40, 40, 1, 1]$, and the random errors were generated from the standard normal distribution. We generated 200 data points as the training data. For the test data, the first set of test data are simulated in the same way as the training except with $\mu_{\text{test1}} = (40, 150, 105, 0, 0)'$, and the second set of test data has $\mu_{\text{test2}} = (70, 150, 105, 0, 0)'$. We generated 30 data points for test set 1 and test set 2.

Using Conn, Johnson, and Boveng (2015), gIVH can be calculated conditioned on a fitted GAM model. So we first fit a GAM model, and use the delta method (Dorfman 1938) to calculate the variance-covariance matrix on the response scale (call this V_{obs}). Then we solved for $\max(\text{var}(\lambda_\mathbf{S}))$ as the maximum diagonal element of V_{obs}. Then we make predictions on the test dataset, and calculate the variance-covariance on the response scale for all prediction locations. Then, for all $i \in \mathbf{S}_0$, we have $\text{var}(\lambda_i)$ and Table 3.

Table 3 gIVH example.

Display Table

From Table 3, test set 1 is within gIVH, since the maximum in the range is still less than $\max(\text{var}(\lambda_\mathbf{S}))$. On the other hand, the majority of the test set 2 are outside the hull. Figure 2 shows the response $\mathbf{y}$ versus x_1, where x_1 is the first column of the design matrix. The test set 1 is more "closer" to training data than test set 2, therefore, test set 1 is within the hull while the test set 2 is not.

Fig. 2 gIVH y versus x_1.

Display full size

4 Prediction Using Random-Effects Extrapolation (PURE)

Suppose we have a set of training data and test data as in (1) and (2). Let π_k denote the percentage of the population that comes from the population k, and $\sum_{k=1}^K \pi_k = 1$. If $K = 2$, we have π_1 percent of the population comes from the minority and the rest $1 - \pi_1$ population comes from the majority. We define the following relevant features:

1. Extreme data: This is the test dataset which may or may not be outside of range of the training data. Both cases can be handled here.

2. Majority data: Notationally, we can concatenate all observations in the full training data as $\mathcal{L} = \{(x_l, y_l); l = 1, \ldots, (n_1 + n_2 + \cdots + n_m)\}$. Then define the majority dataset as those further away from the test data. Let $\ddagger$ denotes the majority, we have a distance measure $d_\ddagger = |\text{median}(\text{var}(\lambda_\ddagger)) \text{-} \max(\text{var}(\lambda_\mathbf{S}))|$ where $\text{var}(\lambda_\ddagger) > \max(\text{var}(\lambda_\mathbf{S}))$ and $\lambda_\ddagger$ denotes the λ that calculated from the majority data. Similarly, $d_\dagger$ denotes the distance measure for the minority data and $d_\ddagger > d_\dagger$. Therefore:

$$\mathcal{L}^\ddagger = \{(x_l, y_l) | d_\ddagger > d_\dagger\}.$$

The original groupings are maintained so the majority data can be re-expressed according to the groupings.

1. Minority data: This portion of the data that is the complement of the majority data. This is found by a minority data decision rule to be described.

$$\mathcal{L}^\dagger = \{(x_l, y_l) | d_\ddagger \le d_\dagger\}.$$

Again, the original groupings are maintained so the minority data can be re-expressed according to these groupings.

1. Regrouped minority data $\mathcal{L}_R^\dagger$: For this, we take the minority data and regroup it according to a hierarchical clustering algorithm with respect to the responses $\mathbf{y}$ resulting in $m_r = m$ groupings with potentially revised memberships.

The full PURE algorithm can then be detailed as follows:

PURE prediction algorithm

1. Fit a Generalized Additive Model (GAM) to the training data using the augmented design matrix where $X_{\text{aug}} = [XZ]$. Then use the delta method to construct the gIVH based on this training set. The GAM is not implicitly necessary and can be replaced by a linear model fit however, the additional flexibility provided by the GAM can capture non-linearities which may otherwise be missed.

2. Use the delta method to estimate variance for predictions for the test data observations; if a prediction is inside the gIVH, use the regular CMMP approach; if a prediction is outside the gIVH, then go to Step 3.

3. Cluster the training data X_{aug} using model-based clustering (Fraley and Raftery 2002), and select the cluster that is closest to the test data using the minority data decision rule described in the next section. Designate this as the training minority data.

4. Regroup the training minority data based on the responses $\mathbf{y}$ using bottom-up hierarchical clustering with the number of groups $m_r = m$ fixed as the original

training data.

5. Fit a regrouped linear mixed-model based on the regrouped minority data and then apply CMMP to obtain the PURE mixed-effects projection for the new test observations denoted $\hat{\theta}_{n,\tilde{\gamma}_X}$.

For step 4, we use a bottom-up hierarchical clustering with average linkage to cluster the training data with respect to the response $\boldsymbol{y}$. The number of groups is set to be the same as in the original training data. Regrouping thus reorganizes the patterns within groups. Its role and effect will be explored later under an asymptotic analysis.

4.1 Minority Data Decision Rule

As previously explained, Conn, Johnson, and Boveng (2015) defined a generalized independent variable hull (gIVH) as a set of all predicted locations $\boldsymbol{S}_0$ for which (9) is true. In PURE, the way to select the minority data is through gIVH. Since the test data is considered extreme data that is outside of the range of the training data, we are not expecting that the test data will fall inside any one of the K gIVHs constructed from K groups of the training data. In applications, if this happens, then the test data is not as "extreme" as we thought, we certainly can use the algorithm to make the prediction. However, the larger gains for our new method happen when the test data is not inside any K gIVHs. In this case, we can always find the sub-group of the training data that is the closest to the test data and designate this as our minority training group.

For K groups of the training data, on the boundary of gIVH, we have the maximum prediction variance of the training sets, which denotes as $\max(\mathrm{var}(\lambda_S))_k, 1 \le k \le K$. For the test data, we have $\mathrm{var}(\lambda_i)_k$ for $1 \le i \le n_{\mathrm{new}}$. Note λ_i corresponds to the mean prediction at i for the test data, and this prediction is different based on different groups of the training data. For each k, we have:

$$d_k = |\mathrm{median}(\mathrm{var}(\lambda_i))_k - \max(\mathrm{var}(\lambda_S))_k|,$$

(12) where $\mathrm{var}(\lambda_i) > \max(\mathrm{var}(\lambda_S))$ for each k, and $1 \le k \le K$. We use $\mathrm{median}(\mathrm{var}(\lambda_i))$ to quantify the overall distance of the test data as a whole to the training data. Among K choices of d_k, the minority data index $\tilde{k}$ is

$$\tilde{k} = \arg\min_k d_k.$$

(13)

Therefore, the test data is assumed to be the closest to the $\tilde{k}$ th training set. The initial K groups in the training data are first identified and fed into the minority data decision rule. Cluster identification is done using model-based clustering in order to stay consistent with the framework 3. In particular, the method of Fraley and Raftery (2002) can be used with the total number of clusters K being tuned by using a Bayesian information criterion (BIC).

In PURE, gIVH is calculated conditioned on a fitted GAM model. In particular, the variance-covariance matrix on the response scale is denoted as V_{obs}. Let $v_{\max}^{(k)}$ equal the maximum diagonal element of V_{obs} for the kth training set. Then we can calculate the variance-covariance on the response scale for all test data locations (call this V_{pred}), and let $V_{\mathrm{pred}}^{\mathrm{diag}}$ denote all the diagonal elements of V_{pred}. We can calculate the distance d as

$$d_k = \left|\mathrm{median}(V_{\mathrm{pred}}^{\mathrm{diag}}) - v_{\max}^{(k)}\right|,$$

(14) where all elements in $V_{\mathrm{pred}}^{\mathrm{diag}}$ are larger than $v_{\max}$ for each k. Here (14) is one frequentist approach of (12), and $\tilde{k}$ is the same as in (13). There is also a Bayesian approach for constructing gIVH, the details of which are described in Conn, Johnson, and Boveng (2015). Finally, the set of training data that has $d_{\tilde{k}}$ is the closest to the test data since its $v_{\max}$ has the minimum distance from the median of $V_{\mathrm{pred}}^{\mathrm{diag}}$. Note (12) and (14) are useful when the test data are outside the hull and we want to see how "far" away they're outside the hull. If all the test data are within the hull, then all $V_{\mathrm{pred}}^{\mathrm{diag}}$ are less than $v_{\max}$, and d_k is not defined in this domain.

In this sense, $d_{\tilde{k}}$ is a quantitative measurement of the distance for the $\tilde{k}$ th population to the test data. Therefore, if the d_ks are very close to each other, the distances of the test data to each of the population are close. However, if $d_{\tilde{k}}$ is smaller than other d_k, then a minority dataset can be identified. Therefore, it is important to examine whether $d_{\tilde{k}}$ is truly smaller than the rest. To do so, we order the d_ks from the smallest to the largest denoted as $d_{(1)}, d_{(2)}, \ldots, d_{(K)}$. Then we require that $d_{(2)} - d_{(1)}$ should be larger than $5\% \times d_{(1)}$. This establishes a threshold for a minority data decision rule. If the threshold is not satisfied, these two datasets are combined and we perform the analysis again. If K equals to 2, this means we only have two datasets—the majority and the minority. The difference $\max(d_1, d_2)$-$\min(d_1, d_2)$ should then be larger than $5\% \times \min(d_1, d_2)$.

5 Simulation Studies

Several simulation studies were carried out to investigate the performance of our new method in comparison with some other competitor approaches. The methods we used for comparison were:

- RP: The least square estimator(RP) method prediction when the grouping information is not known where we only use the fixed effects covariates for prediction.
- CMMP: The original classified mixed model prediction method.
- Lmer*: The method Lmer* uses the random effects for the prediction by assuming the grouping information for the test data observations is known. This method is included comparison purposes, and cannot performed in the real applications if the grouping information for the test data is unknown.
- Regrouped-Lmer*: The same as the above Lmer* method but using the regrouped membership for training.
- Regrouped-CMMP*: The same as the above CMMP method but using the regrouped membership for training.
- PURE: The proposed new method Prediction Using Random-effects Extrapolation using minority data only.
- ULMM: unified linear mixed model (ULMM).
- Random forests and gradient boosting.
- Test data response mean: As indicated in Jiang et al. (2018), $\bar{y}_n$ is another natural estimator that could be used but has limitations. When n_{new} is large, we expect the performance to be similar to CMMP, otherwise, CMMP will be preferable because it borrows strength from other sources through the model.

The ULMM is what we term the *unified linear mixed model (ULMM)*. This fits the model in (1) to the training data. Now suppose a unified LMM is $y_{ij} = w_{ij}'\boldsymbol{\beta} + \alpha_i + \epsilon_{ij}$ where i represents the group for the random effects, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, $w_{ij}' = (x_{ij}', 0)$ if x_{ij} belongs to majority, and $w_{ij} = (0, x_{ij}')$ if x_{ij} belongs to minority. The same random effect group may include both majority and minority individuals. Specifically, if x_{ij} belongs to majority, we have $y_{ij} = x_{ij}'\boldsymbol{\beta}_1 + \alpha_i + \epsilon_{ij}$; and if x_{ij} belongs to minority, we have $y_{ij} = x_{ij}'\boldsymbol{\beta}_2 + \alpha_i + \epsilon_{ij}$.

The predictors have been extended from 6 to 11 (the predictors do not double since one predictor is the intercept, so we have $6 \times 2 - 1 = 11$). If we want to make predictions based on this model, then the test data should be expanded as well. For the test data, $\boldsymbol{\beta}_n$ does not equal either $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_2$. Therefore, the test data are expanded for both sides—that is, $(x_{ij}', 0)$ and $(0, x_{ij}')$, and we obtain two solutions. The model is fit to the full data only since for the minority data, we cannot fit the extended model (some columns are all zero, so they're linearly dependent). Then CMMP is applied.

We will use the same setup as the example in the Introduction but with different scenarios (indexed by Case numbers). Specifically, we change the $\boldsymbol{\beta}$ vectors as in Table 4.

Table 4 Fixed effects parameters when $G = 20$.

Display Table

Case 1 corresponds to the original example in the Introduction. Case 2 sets $\boldsymbol{\beta}_2$ closer to $\boldsymbol{\beta}_1$ compared with Case 1. Therefore, Case 1 is expected to see more of a minority effect since the difference of majority and minority groups is larger. The methods RP, CMMP, and Lmer using the full data are expected to perform better in Case 2 since the difference between the majority and minority groups is smaller than in Case 1. Thus, these models are less "misspecified" compared with Case 1. Case 3 examines the situation when we have the same $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, but a more extreme $\boldsymbol{\beta}_n$.We expect the minority effect to be eliminated. We can, however, expect to observe the regrouping effects more clearly in Case 3. Again, RP, CMMP, and Lmer with the full data are expected to perform better than in Case 1 since the difference of the majority and minority groups gone. Case 4 shows an example where the PURE method (and other methods) would be expected to break down since the $\boldsymbol{\beta}_n$ is quite different. For Case 4, $x_n'\boldsymbol{\beta}_n$ is much larger compared with Case 1 producing much larger response values. This can be observed by evaluating a t-test of the extreme responses and the training data responses. The mean t statistic over 100 sets of training and test is −36.68. For Case 1, 2, and 3, the mean t statistics are -17.31, -21.91, and -17.31, respectively. Case 5 reflects the situation where the test data does not have extreme parameters but instead shares the same regression parameter with the minority data. If some part of the training data relationship is representative of the test data relationship as in Case 5, we would expect that RP, CMMP, and Lmer will work reasonably well and PURE may be not needed. Case 6 is the case where all parameters are equal. It is expected the minority data models will lose efficiency because of the reduced sample size effects, and there is no reason to use PURE in this case.

For the ULMM in the results tables, the column "Average" gives the prediction solution, since the test data is independent of both the majority and the minority and in practice all the true parameters are unknown. Therefore, the solution is obtained by expanding the test data as the majority, then as the minority, and finally, take the average of the two. For all cases, 100 realizations of the simulation were done and the fixed effects design matrix remained constant throughout. We report the average empirical test set mean squared prediction error (MSPE), it's the standard error and the percent decrease in MSPE relative to RP. Tables 5–7 present the results for Cases 1–3 and results tables for Cases 4–6 are presented in the supplementary file. The results can be summarized as follows:

Table 5 Summary of MSPE simulation results for prediction of mixed effect for Case 1.

Display Table

Table 6 Summary of MSPE simulation results for prediction of mixed effect for Case 2.

Display Table

Table 7 Summary of MSPE simulation results for prediction of mixed effect for Case 3.

Display Table

1. For Cases 1, 2, and 3, the full data models and minority data (as estimated by gIVH) models are really mixtures of different linear regressions. Hence, random forests and gradient boosting perform better than RP.
2. For Case 4, the extreme data is much further away from the training data and all methods degrade significantly.
3. For Cases 1 and 2, the minority effect is observed with lower MSPEs based on RP than using the full data.
4. Regrouping effects are observed for Cases 1, 2, and 3 with breakdown happening in Case 4.
5. For Cases 1, 2, and 3, regrouped CMMP and PURE behave very similarly indicating that regrouping has a larger relative effect on reducing MSPE than does also using the minority data regrouping. However, PURE still wins indicating that additional gains are still possible induced by the minority data effect.
6. If the training data follows the same relationship between the majority and minority groups, that is $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ as in Case 3, the benefit of using PURE relative to Regrouped CMMP using full data decreases. The effect of regrouping also decreases when $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ where the full data may contain more information.
7. For ULMM method, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the same for Case 3 and Case 6, using ULMM only doubles the number of parameters, so this unified misspecified model does not have advantages. For these two cases, the "Average" result is very close to the solutions from the previous two columns(left extension and right extension). The three columns of ULMM give similar solutions.
8. PURE performs better than ULMM for Cases 1, 2, 3, and 5. The ULMM has nearly twice the parameters, so there is a tradeoff of a larger sample size for a more complex model. When it comes to prediction, it matters which prediction error contributes more, the majority or minority. The true model ULMM accounts for these errors. For the PURE, however, we move closer to the test data, so the error is reduced.
9. Using the test data response mean as an estimator generally does not perform as well as regrouped CMMP or PURE. The exception to this is in Case 4 (supplementary file) where the test data response mean performs well because the test data is in fact much further away from the training data and hence the training data is of limited value.

6 Asymptotic Analysis

6.1 Misspecified LMM

It is assumed that $\boldsymbol{y}^{(k)}$ follow a misspecified mixed model from (1). Our goal is to show the bias with PURE is smaller than the bias induced by fitting the standard linear mixed model if (i) the prediction for new data are outside the range of the training data and (ii) the training data has a minority group that is closer to the new data. In order to compare the bias, we first study the asymptotic behavior of $\hat{\boldsymbol{\beta}}$ under the misspecified LMM. The reason we call it misspecified LMM is because our setting assumes the training data has at least two populations: majority population and minority population. To simplify the asymptotic analysis, we focus on $K = 2$ where the training data only composed of two populations. For $K = 2$, the majority population is the population that is assumed to have a larger sample size, so we have:

$$\boldsymbol{y}^{(1)} = X_i^{(1)}\boldsymbol{\beta}_1 + Z_i^{(1)}\boldsymbol{b}_i + \epsilon_i,$$

(15) where $i = 1, \ldots, m$. From this population, assume a total sample size is $n^{(1)}$.

The minority population is the population that is closer to the test data and this population is detected by gIVH. For the minority model (assume $k = 2$ is the minority), and we have:

$$\boldsymbol{y}^{(2)} = X_i^{(2)}\boldsymbol{\beta}_2 + Z_i^{(2)}\boldsymbol{b}_i + \epsilon_i,$$

(16) where $i = 1, \ldots, m$. From this population, assume a total sample size is $n^{(2)}$. Note that m is the same for the majority population and the minority population. That means in each of the m groups we have a mixture of majority and minority data.

In the matrix format, let $\boldsymbol{y} = (y_1', \ldots, y_m')' \in \mathbb{R}^n$, $Z = \mathrm{diag}[Z_1, \ldots, Z_m] \in \mathbb{R}^{n \times (mq)}$, $\mathcal{G} = \mathrm{diag}[G, \ldots, G] \in \mathbb{R}^{mq \times (mq)}$, $R = \mathrm{diag}[R_1, \ldots, R_m] \in \mathbb{R}^{n \times n}$, and w be a $n \times n$ matrix with off-diagonal elements equal to zero, and the diagonal elements equal 1 if the observation belongs to the majority, equals 0 if the observation belongs to the minority. Then we have

$$\boldsymbol{y} \sim N(wX\boldsymbol{\beta}_1 + (I - w)X\boldsymbol{\beta}_2, V),$$

(17) where $V = Z'\mathcal{G}Z + R$ and w as defined below.

Let X_i be the matrix consisting of $X_i^{(1)}$ and $X_i^{(2)}$, and let X denote the combined fixed design matrix. Note in each group of X, we have some rows belong to the majority and some rows belong to the minority datasets. Then Z has the similar construction. Now we use a w matrix to "pick up" the observations come from the majority and the minority. This w matrix is a $n \times n$ matrix. So wX pick up all the observations in the majority. Note that w has only diagonal elements, and if the data come from the majority, the diagonal element equals 1, otherwise it equals 0. As a result wX is a $n \times n$ matrix, but the rows belonging to the minority dataset are 0 vectors. Similarly, $(I - w)X$ has nonzero rows for the minority dataset and the rows belonging to the majority dataset are 0 vectors.

The above model is the true model from which the data are generated but when we fit the data we assume there is only one set of $\boldsymbol{\beta}$, so this is where we call the model "misspecified." For the "misspecified" part, we assume $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$. For the misspecified model, we have $\boldsymbol{y} \sim N(X\boldsymbol{\beta}, V)$ since $wX\boldsymbol{\beta} + (I - w)X\boldsymbol{\beta} = X\boldsymbol{\beta}$ when $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

Let $\boldsymbol{\gamma}$ represent the vector of all the variance components in V, and let all the parameters be $\psi = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$. Then the log-likelihood function under misspecified LMM is as follows,

$$\ell(\boldsymbol{y}, \psi) = c - \frac{1}{2}\log\left(|V|\right) - \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})'V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}).$$

(18) where c is some constant. The derivatives of ℓ regarding $\boldsymbol{\beta}$ are:

$$\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{y}, \psi) = X'V^{-1}\boldsymbol{y} - X'V^{-1}X\boldsymbol{\beta},$$

(19)

$$\frac{\partial}{\partial \gamma_j}\ell(\boldsymbol{y}, \psi) = \frac{1}{2}[(\boldsymbol{y} - X\boldsymbol{\beta})'V^{-1}\frac{\partial V}{\partial \gamma_j}V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}) - \mathrm{tr}(V^{-1}\frac{\partial V}{\partial \gamma_j})], j = 1, \ldots, q_*,$$

(20) where γ_j is the jth component of $\boldsymbol{\gamma}$, and it has dimension q_*. Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ denote the MLE. In Jiang (2007), the MLE of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (X'\hat{\mathcal{V}}^{-1}X)^{-1}X'\hat{\mathcal{V}}^{-1}\boldsymbol{y},$$

(21) where $\hat{\mathcal{V}} = V(\hat{\boldsymbol{\gamma}})$. Once we have $\hat{\mathcal{V}}$, the MLE of $\boldsymbol{\beta}$ is given by (21). As for the MLE of $\boldsymbol{\gamma}$, see Jiang (2007) for the detail. We now assume the following regularity conditions.

A1.There exists a unique $\psi_* = (\boldsymbol{\beta}_*', \boldsymbol{\gamma}_*')' \in \Psi^o$, the interior of Ψ, where $M(\psi_*) = \sup_{\psi \in \Psi} M(\psi)$ with $M(\psi) = E(\ell(\boldsymbol{y}, \psi))$.

A2.One can differentiate $E[\ell(\boldsymbol{y}, \psi)]$ with respect to ψ under the expectation, that is, $\frac{\partial}{\partial \psi}E\{\ell(\boldsymbol{y}, \psi)\} = E\{\frac{\partial}{\partial \psi}\ell(\boldsymbol{y}, \psi)\}$.

A3.As $m \to \infty$, we have,

- (i) $\limsup \lambda_{\max}[\frac{1}{m}\sum_{i=1}^m \frac{\partial^2}{\partial\psi\partial\psi'}E\{\ell_i(\boldsymbol{y}_i, \psi_*)\}] < 0$;
- (ii) $\limsup \frac{1}{m}\sum_{i=1}^m E\left\{\frac{\partial}{\partial \psi_r}\ell_i(\boldsymbol{y}_i, \psi_*)\right\}^2 < \infty$;
- (iii) $\frac{1}{m^2}\sum_{i=1}^m E\left\{\frac{\partial^2}{\partial\psi_r\partial\psi_s}\ell_i(\boldsymbol{y}_i, \psi_*)\right\}^2 \to 0$; and,
- (iv) $m^{-3/2}\sum_{i=1}^m E\left\{\sup_{\psi \in S_\rho(\psi_*)}\left|\frac{\partial^3}{\partial\psi_r\partial\psi_s\partial\psi_t}\ell_i(\boldsymbol{y}_i, \psi)\right|\right\} \to 0$ where for some $\rho > 0$ and any $1 \le r, s, t \le q'$ where q' is the dimension of the ψ, and $S_\rho(\psi_*) = \{\psi \in \Psi: |\psi - \psi_*| \le \rho\}$.

Condition A1 and A2 are similar to those in Jiang, Nguyen, and Rao (2011). Condition A1 states ψ_* is identifiable since it is assumed to be the "true" parameter in the sense of maximizing the $M(\cdot)$ function. Condition A2 is a regular condition. Condition A3, (i) is an "information assumption" similar as in Jiang, Nguyen, and Rao (2011). The remainder of A3 are moment conditions. These are mild conditions that are usually satisfied. See Jiang, Nguyen, and Rao (2011) for an example for the illustration of some conditions.

With $M(\psi_*) = E(\ell(\boldsymbol{y}, \psi_*))$, and under (19, 20) and A2 we let

$$E\left\{\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{y}, \psi_*)\right\} = \frac{\partial}{\partial \boldsymbol{\beta}}E\left\{\ell(\boldsymbol{y}, \psi_*)\right\} = \frac{\partial}{\partial \boldsymbol{\beta}}M(\psi_*) = 0,$$

(22) and

$$E\left\{\frac{\partial}{\partial \gamma_j}\ell(\boldsymbol{y}, \psi_*)\right\} = \frac{\partial}{\partial \gamma_j}E\left\{\ell(\boldsymbol{y}, \psi_*)\right\} = \frac{\partial}{\partial \gamma_*}M(\psi_*) = 0,$$

(23) which yield

$$E[\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{y}, \psi_*)] = X'V_*^{-1}(wX\boldsymbol{\beta}_1 + (I - w)X\boldsymbol{\beta}_2) - X'V_*^{-1}X\boldsymbol{\beta}_* = 0,$$

(24) and

$$\begin{aligned} E[\frac{\partial}{\partial \gamma_j}\ell(\boldsymbol{y}, \psi_*)] &= \frac{1}{2}(wX\boldsymbol{\beta}_1 + (I - w)X\boldsymbol{\beta}_2 - X\boldsymbol{\beta}_*)'V_*^{-1}\frac{\partial V}{\partial \gamma_j}V_*^{-1} \\ &\quad (wX\boldsymbol{\beta}_1 + (I - w)X\boldsymbol{\beta}_2 - X\boldsymbol{\beta}_*) \\ &\quad + \frac{1}{2}\mathrm{tr}[V_*^{-1}\frac{\partial V}{\partial \gamma_j}V_*^{-1}(V - V_*)] = 0. \end{aligned}$$

(25)

We now characterize the convergence of $\hat{\boldsymbol{\beta}}$ in the following theorem.

Theorem 1.

Under the conditions A1–A3, there exists with probability tending to one as $m \to \infty$a local maximizer $\hat{\psi}$ of $\ell(\boldsymbol{y}, \psi)$ in a neighborhood of ψ_*, such that $\sqrt{m}(\hat{\psi} - \psi_*) = O_P(1)$. Therefore, focusing on $\boldsymbol{\beta}$, we have $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_* \xrightarrow{P} 0$ as $m \to \infty$, and $\boldsymbol{\beta}_* = \left\{X'V_*^{-1}X\right\}^{-1}X'V_*^{-1}(wX\boldsymbol{\beta}_1 + (I - w)X\boldsymbol{\beta}_2)$, where V_* is given by the solution of (24) and (25).

The proof of Theorem 1 is given in the supplementary materials.

In Theorem 1, $\boldsymbol{\beta}_*$ is some value between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and V_* is given in the appendix. If the full data only has the majority population with $\boldsymbol{\beta}_1$, then the above equation reduces to $\boldsymbol{\beta}_* = (X_1'V_*^{-1}X_1)^{-1}(X_1'V_*^{-1}X_1)\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1$. If the full data only has the minority population with $\boldsymbol{\beta}_2$, then $\boldsymbol{\beta}_* = (X_2'V_*^{-1}X_2)^{-1}(X_2'V_*^{-1}X_2)\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2$. If the full data has both the majority and the minority, then $\boldsymbol{\beta}_*$ is a linear combination of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

An example is given below to further illustrate the limit of $\boldsymbol{\beta}_*$ where a special case with p = 1 is considered. When $p = 1$, we have

$$\begin{aligned} \boldsymbol{\beta}_* &= \left\{X'V_*^{-1}X\right\}^{-1}X'V_*^{-1}(wX\boldsymbol{\beta}_1 + (I - w)X\boldsymbol{\beta}_2) \\ &= (\left\{X'V_*^{-1}X\right\}^{-1}X'V_*^{-1}wX)\boldsymbol{\beta}_1 \\ &\quad + (\left\{X'V_*^{-1}X\right\}^{-1}X'V_*^{-1}(I - w)X)\boldsymbol{\beta}_2. \end{aligned}$$

(26)

Note $(\left\{X'V_*^{-1}X\right\}^{-1}X'V_*^{-1}wX)$ and $(\left\{X'V_*^{-1}X\right\}^{-1}X'V_*^{-1}(I - w)X)$ are now scalars. Let $v_1 = X'V_*^{-1}wX$ and $v_2 = X'V_*^{-1}(I - w)X$, it is clear $v_1 + v_2 = X'V_*^{-1}X$, so $\boldsymbol{\beta}_*$ is the weighted sum of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, and the weights are $\frac{v_1}{v_1 + v_2}$ and $\frac{v_2}{v_1 + v_2}$. Note v_1 involves the data from the majority because of w, and v_2 involves the data from the minority because of $I - w$.

Therefore, in this case, $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_*$ where $\boldsymbol{\beta}_*$ is the weighted sum of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. In real-world applications, the standard approach is to fit one linear mixed model for the training data resulting in the MLE $\hat{\boldsymbol{\beta}}$. The approximate squared bias is $[E(\boldsymbol{\beta}_* - \boldsymbol{\beta}_n)]^2$. Alternatively, by PURE, the approximate squared bias is $[E(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_n)]^2$ after the minority data is correctly selected by gIVH. By design, $\boldsymbol{\beta}_n$ is closer to $\boldsymbol{\beta}_2$, hence, $[E(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_n)]^2 \le [E(\boldsymbol{\beta}_* - \boldsymbol{\beta}_n)]^2$, since the weighted contribution of $\boldsymbol{\beta}_1$ in $\boldsymbol{\beta}_*$ pulls it away from $\boldsymbol{\beta}_n$.

6.2 The Regrouping Effect

The next theorem details what we term the regrouping effect and how this leads to lower MSPE for PURE as compared to the CMMP prediction from the misspecified LMM.

Conditions B1-B6 are defined as follows. Conditions B1-B4 are from Jiang et al. (2018) supplementary materials (from now on called JRFN suppl (2018)) for consistency of the usual CMMP estimator.

B1. $\boldsymbol{b}_i, 1 \le i \le m, \boldsymbol{b}_{\mathrm{new}}$ are independent with the ϵ_{ij}'s where $\boldsymbol{b}_{\mathrm{new}}$ refers to the test data random effects; $\boldsymbol{b}_{\mathrm{new}}$ is independent with the training data when it is a new random effect, and ϵ_n is independent with all of the $\boldsymbol{b}_i$'s and ϵ_{ij}'s.

B2. $|x_{ij}|, 1 \le i \le m, 1 \le j \le n_i$ and x_n are bounded.

B3. $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_*$ in L^2, where $\boldsymbol{\beta}_*$ is the same $\boldsymbol{\beta}_*$ from Theorem 1.

B4. $m \to \infty$, $(\log m)^{2v}/n_{\min} \to 0$, where v is a positive constant, $n_{\min} = \min_{1 \le i \le m} n_i$, and $n_{\mathrm{new}} \to \infty$, where n_{new} is the size of the test data.

B5. For any $1 \le i \le m$, the function $h_i(\rho) = (x_n - \bar{x}_{i\cdot})'\boldsymbol{\beta}_*$, where $\boldsymbol{\beta}_* = (1 - \rho)\boldsymbol{\beta}_1 + \rho\boldsymbol{\beta}_2$ and $\rho \in [0, 1]$, is nondecreasing in ρ. Also, for $1 \le i \le m$, we have, $x_n'\boldsymbol{\beta}_2 \le x_n'\boldsymbol{\beta}_n$, and $\bar{x}_{i\cdot}'\boldsymbol{\beta}_1 < \bar{x}_{i\cdot}'\boldsymbol{\beta}_2$.

B6. After regrouping, for each minority group i, $|E(\bar{y}_{i\cdot}) - \bar{x}_{i\cdot}'\boldsymbol{\beta}_2| \to 0$ where $\boldsymbol{\beta}_2$ is the minority parameter vector.

Conditions B1–B4 are regularity conditions similar to those in Jiang, Lahiri, and Wan (2002), Das, Jiang, and Rao (2004), and Jiang et al. (2018). Condition B5 indicates the setting of our extrapolation problem. The new observation is more extreme than the observations in group i. This means that x_n is "larger" than $\bar{x}_{i\cdot}$; therefore, the difference $x_n - \bar{x}_{i\cdot}$ is "positive." Then, as ρ increases from 0 to 1, $\boldsymbol{\beta}_*$ "increases" from $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_2$. Therefore, the function $h_i(\rho)$ increases as $\boldsymbol{\beta}_*$ increases. For the second part of Condition B5, it indicates that the new observations are closer to the minority data than the majority data. Condition B6 says that the regrouping asymptotically preserves the correct mean function, which seems to be a reasonable assumption. Note that this assumption is actually weaker than assuming that the regrouping consistently identifies the correct minority groups (i.e., one can have incorrect grouping but still get the mean right).

We can then state the following theorem regarding the regrouping effect:

Theorem 2.

Under the conditions B1–B6, when $n_{\min} \to \infty$ we have

$$\limsup E(\hat{\theta}_{n,\tilde{\gamma}_X} - \theta_n)^2 \le \limsup E(\hat{\theta}_{n,\tilde{I}} - \theta_n)^2,$$

where $\hat{\theta}_{n,\tilde{\gamma}_X}$ is the CMMP estimator for the regrouped data, and $\hat{\theta}_{n,\tilde{I}}$ is the CMMP estimator for the original data.

The proof of Theorem 2 is detailed in the supplementary materials.

7 Real-Data Example: Predicting Age of Smoking Onset from the National Longitudinal Mortality Study

The National Longitudinal Mortality Study (NLMS) (https://www.census.gov/did/www/nlms/about/projectDescription.html)[1] consists of a database developed for the purpose of studying the effects of demographic and socio-economic characteristics on differentials in U.S. mortality rates. The NLMS is a unique research database in that it is based on a random sample of the non-institutionalized population of the United States. It consists of U.S. Census Bureau data from Current Population Surveys, Annual Social and Economic Supplements and a subset of the 1980 Census combined with death certificate information to identify mortality status and cause of death. The study currently consists of approximately 3.8 million records with over 550,000 identified mortality cases.

We used the tobacco-use data from the NLMS, which has a maximum follow-up time of 5 years and consists of data centered around the year 2000. This gave us 192,764 observations where the variables age at onset of smoking (agesmk), race, age, sex, marital status, weight, number in the household, and follow-up time have no missing values. We were interested in trying to predict agesmk from the other predictors. We randomly sampled 2% observations from the data resulting in a training dataset with the following distribution by race:

We will start out by entertaining the idea that the non-White subjects could be considered an extreme test set for which predictions built on training data from White and Black individuals would be considered an extrapolation (Table 8).

Table 8 NLMS cohort broken down by race.

Download CSV Display Table

To generate groups of test observations from which we will predict the associated mixed effect as in Jiang et al. (2018), we need to first estimate the grouping structure in the test data. Thus, we clustered the test responses using hierarchical clustering and estimated the number of groups using the cross-validation method of Wang (2010), we estimate the number of test data groups to be $\bar{m} = 2$. We then constructed groups of five test observations each with a fixed x_n value resulting in a total of 290 observations for the test data. Specifically, we were guided by the approach used by Jiang et al. (2018) with a twist. After we fit a linear mixed model to the test data, we then generated additional mixed effects as in Jiang et al. (2018). However, the random effects were not matched to the training data random effects but instead, the estimated test data random effects were attached to the newly generated test data fixed effects. Hence, for the real data analysis, we are in the "unmatched" setting and our methodology is trying to find the closest value to the test data random effects from amongst the training data groups. For the full training data, we estimated $\bar{m} = 20$ by clustering the responses using hierarchical clustering and estimated the number of clusters again using the cross-validation method of Wang (2010).

Using the gIVH approach, majority and minority datasets were separated with 2218 and 1517 observations, respectively. Table 9 shows the empirical mean squared prediction errors for predicting the test set mixed effects. It's clear that the minority effect is present but less impactful than the regrouping. PURE predictions do a bit better than regrouped CMMP predictions using the full data or the minority data with an empirical MSPE of 0.02638 versus 0.05956. This is because the minority data is somewhat closer to the non-White test observations.

Table 9 NLMS MSPE(SD) for prediction of mixed effect.

Download CSV Display Table

Figure 3, shows boxplots of the variance for prediction locations based on the full training data, majority and minority datasets (as defined by our gIVH algorithm) and the gIVH boundaries as indicated by the red points for the non-White test observations. It's quite clear that indeed the non-White subjects are well outside the range of the other datasets and that the minority dataset is indeed closer to the test data than the majority dataset is.

Fig. 3 Boxplots of GIVH.

Display full size

7.1 Data-based Identification of an Extreme Test Set

One might wonder if a data-based strategy for identifying an extreme dataset could be used instead of using the non-White individuals apriori (even though they were well outside of the gIVH). So, we combined White, Black, and non-White individuals as the full data and did a principal components analysis of the variance-covariance matrix of the predictors. Figure 4 shows the first principal component (PC1) plotted against the second principal component (PC2). It seems a large number of observations are clustered together with PC1 values less than 0. Thus, it might be reasonable to consider those with $PC1 > 0$ as an extreme test dataset. The racial composition for new training and test datasets defined in this way are shown in Tables 10 and 11, respectively.

Fig. 4 Scatterplot of PC1 versus PC2 for all of the data.

Display full size

Table 10 NLMS PURE population for $PC1 > 0$ as test data.

Display Table

Table 11 NLMS PURE population for $PC1 < 0$ as training data.

Display Table

We use the larger set as the training set and the smaller set as the test set. The full test data has 202 observations, and $\widehat{m}$ is estimated to be 2. After once again generating five test responses per x_n, we have 1010 total test observations. For the majority and minority, we set $\widehat{m}$ to be 20. The hierarchical clustering algorithm and gIVH is used to decide on majority or minority data and the data separated as two clusters with 1345 and 2246 observations for each respectively. Figure 5 shows the variance-covariance for all prediction locations (predictions for the test data), which indicates the separation of the training data into majority and minority based on PCA is not as clear as it was before when the non-whites were considered an extreme dataset.

Fig. 5 Boxplots of gIVH with data-based estimation of extreme dataset.

Display full size

Figure 6 shows a zoomed-in version of Figure 5 and it's quite clear that the test data is now at times within the gIVH hull of the minority data. Figure 7 shows another view of this plotting the scaled variance on one y-axis and the response log(agesmk) on the other y-axis against the predictor log(age). The dashed line at the bottom is the gIVH boundary based on the minority data. The individual data points on the plot are the raw data values for the minority data. The jagged line is the gIVH for individual test data points. It's clear that a few of them are below the gIVH boundary and thus within the gIVH hull of the minority data.

Fig. 6 Boxplots of GIVH (Enlarged for detail).

Display full size

Fig. 7 gIVH of minority(two y-scales).

Display full size

Fig. 8 Scatterplot of random effects for CMMP versus PURE.

Display full size

Table 12 shows the MSPE results from this analysis. Clearly, massive gains in prediction accuracy are realized when using PURE with over 99% reduction in MSPE relative to RP. However, similar trends are also seen for other regrouping approaches.

Table 12 NLMS MSPE(SD) for prediction of mixed effect(Minority data by PCA)

Download CSV Display Table

Remark.

We found that the random sampling of 2% of the observations from the full dataset to create training data can sometimes result in majority and minority datasets being less well-separated. In this case, standard CMMP typically performs the best.

7.2 Comparison of Classified Random Effects—Evidence of Extrapolation

We further explored the degree of empirical evidence for random effect extrapolation using PURE based on the PCA analysis approach. Figure 8 plots the absolute values of the classified random effects of CMMP against PURE. Figure 9 plots the same thing using regrouped CMMP versus PURE. Once again, the random effect extrapolation is clearly evident due to regrouping and PURE. In fact, the regrouped CMMP using the full training data often produced classified random effects with larger magnitudes than PURE.

Fig. 9 Scatterplot of random effects for Regrouped CMMP versus PURE.

Display full size

8 Discussion

In this article we developed a novel prediction method for generating more accurate model projections which is inherently difficult due to a lack of observed training data. As a result, accuracy of such predictions with traditional methods (and even some more modern machine learning approaches) is very low. We have shown that for certain types of model structure, our PURE method can make use of a novel statistical construct we term an extrapolated random effect, which as in the original CMMP methodology, acts like a *sponge* to capture what is uncaptured by the model. However in PURE, the role of the classified random effects is even more important. The extrapolated random effects amount to using an extrapolated model for the variance structure which acknowledges the additional uncertainty inherent to extrapolation problems. Importantly, this can be derived with a simple regrouping and relocating strategy as we showed.

Supplemental material

Supplemental Material

Download Zip (195.7 MB) **Supplemental Material**

Disclosure Statement

The author(s) report there are no competing interests to declare.

Additional information

Funding

JSR was partially supported by NSF grants DMS-1915976 and DMS-2210208 and JJ was partially supported by NSF grants DMS-1914465 and DMS-2210569.

Notes

1 This article uses data obtained from the public-use file of the National Longitudinal Mortality Study. The file is provided to persons interested in research by the U.S. Census Bureau. The views expressed in this article are those of the authors and do not necessarily reflect the views of the National Longitudinal Mortality Study, the Bureau of the Census, or the project sponsors: the National Heart, Lung, and Blood Institute, the National Cancer Institute, the National Institute on Aging, and the National Center for Health Statistics. Please see https://www.census.gov/did/www/nlms/about/projectDescription.html.

References

1. Armstrong, J. S. (1984), "Forecasting by Extrapolation: Conclusions from 25 Years of Research," *Interfaces*, 14, 52–66. DOI: 10.1287/inte.14.6.52.

   Web of Science ® | Google Scholar | Get it at UC

2. Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Assoccation*, 83, 28–36. DOI: 10.1080/01621459.1988.10478561.

   Web of Science ® | Google Scholar | Get it at UC

3. Chen, Y. Q., and Cheng, S. (2006), "Linear Life Expectancy Regression with Censored Data," *Biometrika*, 93, 303–313. DOI: 10.1093/biomet/93.2.303.

   Web of Science ® | Google Scholar | Get it at UC

4. Conn, P. B., Johnson, D. S., and Boveng, P. L. (2015), "On Extrapolating Past the Range of Observed Data When Making Statistical Predictions in Ecology," *PloS One*, 10, e0141416. DOI: 10.1371/journal.pone.0141416.

   PubMed | Web of Science ® | Google Scholar | Get it at UC

5. Cook, R. D. (1979), "Influential Observations in Linear Regression," *Journal of the American Statistical Association*, 74, 169–174. DOI: 10.1080/01621459.1979.10481634.

Web of Science ® | Google Scholar | Get it at UC

6. Das, K., Jiang, J., and Rao, J. N. K. (2004), "Mean squared error of empirical predictor," *Technical Report*.

   Google Scholar | Get it at UC

7. Dorfman, R. (1938), "A Note on the Delta-Method for Finding Variance Formulae," *The Biometric Bulletin*, 1, 129–137.

   Google Scholar | Get it at UC

8. Ehrenberg, A. S. C., and Bound, J. A. (1993), "Predicability and Prediction," *JRSS-A*, 156, 167–206. DOI: 10.2307/2982727.

   Web of Science ® | Google Scholar | Get it at UC

9. Fine, T. (1970), "Extrapolation When Very Little is Known About the Source," *Information and Control*, 16, 331–359. DOI: 10.1016/S0019-9958(70)90177-4.

   Google Scholar | Get it at UC

10. Fraley, C., and Raftery, A. E. (2002), "Model-based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. DOI: 10.1198/016214502760047131.

    Web of Science ® | Google Scholar | Get it at UC

11. Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, CRC Press.

    Google Scholar | Get it at UC

12. Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and their Applications*, New York: Springer.

    Google Scholar | Get it at UC

13. Jiang, J., and Lahiri, P. (2006), "Mixed Model Prediction and Small Area Estimation," *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 15, 1–96. Available at https://EconPapers.repec.org/RePEc:spr:testjl:v:15:y:2006:i:1:p:1-96.

    Google Scholar | Get it at UC

14. Jiang, J., Lahiri, P., and Wan, S. M. (2002), "A Unified Jackknife Theory for Empirical Best Prediction with M-estimation," *The Annals of Statistics*, 30, 1782–1810. DOI: 10.1214/aos/1043351257.

    Web of Science ® | Google Scholar | Get it at UC

15. Jiang, J., Nguyen, T., and Rao, J. S. (2011), "Best Predictive Small Area Estimation," *Journal of the American Statistical Association*, 106, 732–745. DOI: 10.1198/jasa.2011.tm10221.

    Web of Science ® | Google Scholar | Get it at UC

16. Jiang, J., Rao, J. S., Fan, J., and Nguyen, T. (2018), "Classified Mixed Model Prediction," *Journal of the American Statistical Association*, 113, 269–279. DOI: 10.1080/01621459.2016.1246367.

    Web of Science ® | Google Scholar | Get it at UC

17. Rao, J. N. K., and Molina, I. (2015), *Small Area Estimation*, Wiley.

    Google Scholar | Get it at UC

18. Saei, A., and Chambers, R. (2005a), "Empirical Best Linear Unbiased Prediction for Out of Sample Areas," Working paper m05/03, Southampton Statistical Sciences Research Institute, University of Southampton, UK.

    Google Scholar | Get it at UC

19. Saei, A., and Chambers, R. (2005b), "Out of Sample Estimation for Small Areas Using Area Level Data," Working paper m05/11, Southampton Statistical Sciences Research Institute, University of Southampton, UK.

Google Scholar | Get it at UC

20. Sun, H., Jiang, J., Nguyen, T., and Luan, Y. (2018), "Best Look-Alike Prediction: Another Look at the Bayesian Classifier and Beyond," *Statistics and Probability Letters*, 143, 37–42. DOI: 10.1016/j.spl.2018.07.014.

Web of Science ® | Google Scholar | Get it at UC

21. Walz, G. (1991), "The History of Extrapolation Methods in Numerical Analysis," Technical Report, Universitat Mannheim, 130.

Google Scholar | Get it at UC

22. Wang, J. (2010), "Consistent Selection of the Number of Clusters via Crossvalidation," *Biometrika*, 97, 893–904. DOI: 10.1093/biomet/asq061.

Web of Science ® | Google Scholar | Get it at UC

Download PDF

Information for

Authors
R&D professionals
Editors
Librarians
Societies

Open access

Overview
Open journals
Open Select
Dove Medical Press
F1000Research

Opportunities

Reprints and e-prints
Advertising solutions
Accelerated publication
Corporate access solutions

Help and information

Help and contact
Newsroom
All journals
Books

Keep up to date

Register to receive personalised research and resources by email

Sign me up

Copyright © 2025 Informa UK Limited Privacy policy Cookies Terms & conditions Accessibility

Registered in England & Wales No. 01072954
5 Howick Place | London | SW1P 1WG

Taylor & Francis Group
an informa business