Submitted to Annals of Applied Statistics

A DISTRIBUTION-FREE MIXED-INTEGER OPTIMIZATION APPROACH TO HIERARCHICAL MODELLING OF CLUSTERED AND LONGITUDINAL DATA

BY MADHAV SANKARANARAYANAN[1,a], INTEKHAB HOSSAIN[2,b] AND TOM CHEN[3,c]

[1]*Department of Biostatistics, Harvard T.H.Chan School of Public Health,* [a]*madhav_sankaranarayanan@g.harvard.edu*

[2]*Department of Biostatistics, Harvard T.H.Chan School of Public Health,* [b]*ihossain@g.harvard.edu*

[3]*Department of Population Medicine, Harvard Medical School,* [c]*tchen@hsph.harvard.edu*

We study linear mixed models in settings where only a small subset of clusters depart meaningfully from the population mean. We cast estimation as a mixed-integer optimization problem that imposes ℓ_0 sparsity on random effects while jointly estimating fixed effects, yielding exact zeros for many clusters and interpretable effect estimates for those retained. The formulation admits certified near-optimal solutions at practical scales using modern MIO solvers. Operating characteristics are examined in simulations spanning two regimes: approximately Gaussian random effects and heterogeneous settings in which deviations are concentrated in a few clusters. When the Gaussian specification holds, the proposed estimator performs comparably to Gaussian mixed models for fixed-effect estimation and prediction; when heterogeneity is sparse, it improves estimation accuracy and reduces test error on held-out clusters by focusing attention on atypical units. Prediction is evaluated using a cluster-blocked validation scheme, and we outline a supervised mapping from cluster-level features to predicted random effects to support prediction for new clusters. Two applications demonstrate the approach. Using New York State hospital-acquired infection data for 2019 hysterectomy, the method recovers the hospitals flagged by public reporting while producing a sparse set of hospital effects. In a student performance dataset, it identifies adjusted high- and low-performing students that are not evident from raw scores. A brief protein expression example appears in the Supplement. These results suggest that MIO-based sparsity on random effects is a useful complement to classical mixed-model analysis when the goal is to flag atypical units and improve prediction for new clusters.

1. Introduction. A recurring challenge in applied statistics is identifying organizational units that significantly over- or under-perform. For example, in education researchers routinely use value-added models to compare teachers' contributions to student achievement, estimating teacher-level effects after controlling for student covariates and school assignment policies (Chetty et al., 2014; McCaffrey et al., 2004). Similarly, in healthcare provider benchmarking, hospitals are ranked after risk adjustment based on hospital-level random-effects (Hota et al., 2020). Both problems lead to a hierarchical formulation where clusters (teachers, hospitals) can be treated as random effects, and the goal is to identify the subset with substantively nonzero effects. Historically, subset-selection efforts focused on selecting the most informative *fixed* predictors (i.e., β coefficients) (Bertsimas et al., 2015); later extensions introduced spline models for non-linear fixed effects (Kartal Koc and Bozdogan, 2015) and mixed models to capture cluster structure (Kowal, 2023). With the advent of Generalized Estimating Equations and mixed-effects methodologies (Liang and Zeger, 1986),

Keywords and phrases: clustered data, significant cluster selection, mixed integer programming, random effect ranking.

attention turned from mean structures alone to correlation and clustering. More recently, researchers have begun focusing on inference about clustering itself, including the magnitude and variability of cluster effects (Chen et al., 2020; Oberauer, 2022). Yet in many contexts, practitioners still treat all clusters as potentially active rather than selecting which clusters have substantively non-zero effects.

We formulate *significant cluster selection* (SCS) to identify a sparse subset of clusters whose random effects are substantively nonzero. We cast SCS as a mixed-integer optimization (MIO) problem with random effects as selection variables, yielding interpretable choices of significant clusters with guarantees of near-optimality for the fitted model. Conceptually, this is the clustered analogue of best-subset selection for fixed effects, but targeted to random-effect heterogeneity (e.g., hospitals with genuine departures after risk adjustment). We assess tractability in applied terms, solving realistically sized instances to small, certified optimality gaps within practical time, rather than by asymptotic polynomial criteria. This optimization view provides transparent sparsity control, clear decision rules for inclusion/exclusion, and a bridge from estimation to action in ranking applications.

Prior work on linear mixed models has largely emphasized robustness and distributional flexibility for random effects by broadening the modeling of $\boldsymbol{\gamma}_k$ to stabilize prediction and inference rather than to induce sparsity (Lin and Lee, 2008; Lin, 2008; Pinheiro et al., 2001; Verbeke and Lesaffre, 1996; Chen and Wang, 2020). In contrast, ℓ_1-type regularization has been proposed for random effects (Geraci and Farcomeni, 2020), but primarily as a shrinkage device; it does not target which clusters should be retained and which should be excluded, i.e., it does not address significant cluster selection (SCS). Moreover, the well-known limitations of LASSO for subset selection of fixed effects carry over: under noise and correlated covariates, ℓ_1 penalties tend to produce many small, biased nonzeros (including spurious ones) and can miss true signals when regularity conditions fail (Greenshtein, 2006; Mazumder et al., 2011; Raskutti et al., 2011; Zhang et al., 2014). For cluster-level decisions (such as ranking or flagging units) this bias-selection trade-off is especially consequential: excessive shrinkage blurs large effects, while diffuse nonzeros obscure which clusters truly deviate. These gaps motivate an approach that explicitly selects clusters rather than merely shrinking them.

The remainder of the paper is organized as follows. Section 2 introduces the linear mixed-effects model, formalizes the sparse random-effects problem at the cluster level, presents the MIO formulation, and details model selection, prediction for new clusters, and implementation. Section 3 reports simulation studies comparing estimation accuracy, predictive performance on held-out clusters, and computational cost against Gaussian and Laplace LMM baselines. Section 4 applies the methodology to New York State hospital-acquired infection data and to a student performance dataset; a brief protein expression example appears in the Supplement. Section 5 discusses practical guidance, limitations, and methodological extensions, with software and data availability provided at the end.

2. Methods.

2.1. *Model framework.* We consider data arising from a clustered or hierarchical structure, where observations are nested within organizational units (e.g., hospitals, schools, or research sites). Let Y_{ki} denote the outcome of the ith observation in cluster k ($k = 1, \cdots, K$, $i = 1, \cdots, n_k$). Let $\mathbf{X}_{ki} \in \mathbb{R}^{P+1}$ and $\mathbf{Z}_{ki} \in \mathbb{R}^{Q+1}$ denote covariate vectors for fixed and random effects, respectively, both including a fixed and a random intercept term. Following the linear mixed-effects model (LMM) of Laird and Ware (1982), we specify

$$Y_{ki} \,|\, \boldsymbol{\gamma}_k = \mathbf{X}_{ki}^\top \boldsymbol{\beta} + \mathbf{Z}_{ki}^\top \boldsymbol{\gamma}_k + \varepsilon_{ki}, \qquad \varepsilon_{ki} \sim \mathcal{N}(0, \sigma_\varepsilon^2), \qquad \boldsymbol{\gamma}_k \sim F_\gamma, \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of population-level (fixed-effect) coefficients, and $\boldsymbol{\gamma}_k$ captures cluster-specific deviations that are typically modeled as independent, mean-zero draws from a distribution F_γ with covariance $\boldsymbol{\Sigma}_\gamma$. This hierarchical structure accommodates within-cluster correlation and allows both fixed and random contributions to the mean response.

Ignoring random effects (that is, neglecting clustering) can distort inference and lead to biased or inefficient estimates in practice (Ntani et al., 2021). In many applications, heterogeneity is sparse: most clusters lie near the population mean while a small subset deviates systematically. This shifts attention from modeling the full distribution F_γ to identifying which clusters depart materially.

2.2. *Significant Cluster Selection (SCS) problem.* We decompose the random effects as $\boldsymbol{\gamma}_k = (\gamma_{k0}, \cdots, \gamma_{kQ})^\top,\ k = 1, \cdots, K$, and define, for each random-effect component $r = 0, \cdots, Q$, the vector $\boldsymbol{\gamma}'_r = (\gamma_{1r}, \cdots, \gamma_{Kr})^\top$ collecting cluster-specific effects across all K clusters. Classical LMM estimation treats all entries of $\boldsymbol{\gamma}'_r$ as potentially nonzero, whereas SCS seeks a sparse representation in which only a subset of clusters contributes to between-cluster variation. In other words, we estimate both the fixed effects $\boldsymbol{\beta}$ and the active set $\mathcal{A}_r = \{\, k : \gamma_{kr} \neq 0 \,\}, |\mathcal{A}_r| = \lambda_r$, where λ_r directly controls the number of clusters allowed to deviate for random-effect component r. The SCS framework generalizes the mixed-effects model to a setting where we solve

$$\min_{\boldsymbol{\beta},\boldsymbol{\Gamma}} \frac{1}{2}\sum_{k=1}^{K} \|\mathbf{Y}_k - \mathbf{X}_k\boldsymbol{\beta} - \mathbf{Z}_k\boldsymbol{\gamma}_k\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\gamma}'_r\|_0 \leq \lambda_r,\ \ r = 0, \cdots, Q, \tag{2}$$

where $\|\cdot\|_0$ counts the number of nonzero elements. The ℓ_0 constraints explicitly enforce sparsity at the cluster level, isolating those units whose random effects are meaningfully nonzero while shrinking the remainder exactly to zero. The parameters λ_r serve as interpretable sparsity controls, each determining how many clusters are permitted to exhibit deviations for a given random-effect term. We note that continuous shrinkage methods such as the LASSO bias all estimates toward zero and can blur the distinction between negligible and substantive cluster effects under correlation; in contrast, the ℓ_0 formulation yields discrete inclusion decisions at the cluster level, thereby producing an interpretable active set $\mathcal{A}_r$.

2.3. *Mixed-Integer Optimization (MIO) formulation.* The optimization problem in (2) involves ℓ_0 constraints that render it nonconvex and combinatorial. Mixed-integer optimization (MIO) provides a natural framework for obtaining globally or near-globally optimal solutions to such best-subset problems while retaining explicit control over sparsity. We follow the formulation and computational strategies of Bertsimas et al. (2015), adapting them to the hierarchical structure of the mixed-effects model.

Let $\mathbf{Y}_k \in \mathbb{R}^{n_k}$, $\mathbf{X}_k \in \mathbb{R}^{n_k \times (P+1)}$, and $\mathbf{Z}_k \in \mathbb{R}^{n_k \times (Q+1)}$ denote the stacked outcome and design matrices for cluster k. Define $\overline{\mathbf{Y}} = [\mathbf{Y}_1^\top : \cdots : \mathbf{Y}_K^\top]^\top$ and similarly $\overline{\mathbf{X}}$ and $\overline{\mathbf{Z}}$. For each random-effect component r, we introduce a binary inclusion vector $\mathbf{s}_r = (s_{1r}, \cdots, s_{Kr})^\top \in \{0,1\}^K$ such that $s_{kr} = 1$ indicates that cluster k is active for component r. The constraint $\sum_{k=1}^K s_{kr} \leq \lambda_r$ enforces the sparsity level from (2). To encode the ℓ_0 restrictions, we link the binary and continuous variables by big-M constraints $-M_r s_{kr} \leq \gamma_{kr} \leq M_r s_{kr},\ k = 1, \cdots, K,\ \ r = 0, \cdots, Q$ where $M_r > 0$ is a sufficiently large constant bounding the feasible range of each γ_{kr}. The resulting MIO formulation is

$$\min_{\boldsymbol{\beta},\boldsymbol{\Gamma},\mathbf{s}} \quad \frac{1}{2}\sum_{k=1}^{K} \|\mathbf{Y}_k - \mathbf{X}_k\boldsymbol{\beta} - \mathbf{Z}_k\boldsymbol{\gamma}_k\|_2^2 + \mu\|\boldsymbol{\beta}\|_2^2, \tag{3}$$

$$\text{s.t.} \quad -M_r s_{kr} \le \gamma_{kr} \le M_r s_{kr}, \quad k = 1, \cdots, K,\ r = 0, \cdots, Q,$$
$$\sum_{k=1}^{K} s_{kr} \le \lambda_r, \quad s_{kr} \in \{0,1\}, \quad r = 0, \cdots, Q.$$

The small ridge penalty $\mu\|\boldsymbol{\beta}\|_2^2$ stabilizes the optimization and ensures numerical convergence without materially affecting sparsity or interpretability. Equation (3) is the clustered analogue of best–subset regression, where the selection units are clusters rather than covariates. This formulation admits a decomposition into an *inner* continuous estimation and an *outer* combinatorial selection. For any fixed pattern of binary inclusion variables $\mathbf{s}$ (i.e., a fixed active set of clusters), the inner problem is a convex ridge–regularized least–squares fit in $(\boldsymbol{\beta}, \boldsymbol{\Gamma})$ with a closed–form solution; substituting that solution yields a profiled objective $c(\mathbf{s})$ that depends only on the selection pattern. The outer problem then searches over feasible $\mathbf{s}$ subject to the sparsity budgets $\sum_k s_{kr} \le \lambda_r$. We solve the outer problem using an cutting-planes scheme (Duran and Grossmann, 1986): linear inequalities are iteratively added to bound the profiled loss $c(\mathbf{s})$ on the $\{0,1\}$ domain, producing a sequence of mixed–integer linear relaxations that converge to the certified near–optimal selection. Implementations are in `Julia` with the `Gurobi` solver (Gurobi Optimization, LLC, 2023). Theoretical properties and convergence guarantees for related formulations appear in Fukushima (1984); Del Pia and Weismantel (2010), and additional justification is provided in the Supplemental Material.

2.4. *Model selection and tuning.* A key hyperparameter in the SCS formulation (2) is the sparsity level λ_r, which determines how many clusters are allowed to deviate for each random-effect component $r = 0, \cdots, Q$. Selecting λ_r balances parsimony and predictive accuracy: overly small values risk omitting genuinely heterogeneous clusters, whereas large values may admit noise-driven deviations. We determine λ_r by grid search over a finite set of candidate values using a cluster-blocked validation scheme. Specifically, clusters are partitioned into training and validation sets so that all observations from a given cluster remain together. For each candidate λ_r, the MIO problem (3) is solved on the training data, and predictive mean squared error (MSE) is evaluated on the held-out clusters. The λ_r that minimizes validation MSE is then selected and used for refitting on the combined training and validation data. Therefore, performance is assessed on genuinely unseen clusters, guarding against optimistic bias due to within-cluster dependence. The ridge coefficient μ serves only to stabilize numerical optimization and can be fixed at a small constant (e.g., $\mu = 10^{-4}$) without materially affecting accuracy or sparsity.

2.5. *Prediction for new clusters.* A central advantage of the SCS framework is its ability to provide interpretable predictions for new or previously unseen clusters. Standard mixed-effects models implicitly assume that all cluster effects arise from a common distribution F_γ, which can be limiting when new clusters differ systematically from those used in model fitting. To address this, we learn an explicit mapping between cluster-level covariates and their estimated random effects, enabling informed prediction of cluster deviations for new data.

Let $\mathbf{X}_k$ denote the covariate matrix for cluster k and $\widehat{\boldsymbol{\gamma}}_k$ the estimated cluster effects obtained from the MIO formulation. Using the training data, we fit a supervised model that predicts $\widehat{\boldsymbol{\gamma}}_k$ from summary features of $\mathbf{X}_k$, written as $f : \mathbf{X}_k \mapsto \widehat{\boldsymbol{\gamma}}_k$. For interpretability, we employ transparent classification or regression algorithms such as Classification and Regression Trees (CART; Breiman et al., 2017) or Optimal Classification Trees (OCT; Bertsimas and Dunn, 2017), which provide simple rule-based partitions linking observable cluster characteristics to their estimated random effects. This mapping allows new observations to be

assigned to existing clusters or to combinations thereof, facilitating out-of-sample prediction.

For a new observation with covariate vector $\mathbf{X}_\nu$, let π_ν denote the predicted distribution over clusters based on the learned mapping. We consider two operational strategies:

- **Hard assignment:** Assign the observation to the most probable cluster and apply its estimated effect,

$$\widehat{Y}_\nu = \mathbf{X}_\nu^\top \widehat{\boldsymbol{\beta}}_{\text{MIO}} + \widehat{\gamma}_{\arg\max(\pi_\nu)}$$

- **Soft assignment:** Compute a weighted average of cluster effects using π_ν as weights,

$$\widehat{Y}_\nu = \mathbf{X}_\nu^\top \widehat{\boldsymbol{\beta}}_{\text{MIO}} + \pi_\nu^\top \widehat{\boldsymbol{\gamma}}$$

The soft-assignment approach is generally more robust when the relationship between covariates and cluster identity is uncertain, as it accounts for predictive uncertainty through the distribution π_ν. Both strategies extend the utility of SCS beyond the original set of clusters, enabling interpretable generalization to new organizational units (e.g., new hospitals, schools, or sites) without assuming exchangeability. In subsequent empirical analyses, we adopt the soft-assignment strategy due to its stability under model misspecification and its ability to weight existing clusters by similarity rather than force discrete membership.

2.6. *Algorithmic implementation and computational details.* The algorithmic pipeline is implemented in `Julia` with `Gurobi` as the mixed-integer optimizer. A schematic is shown in Figure 1.

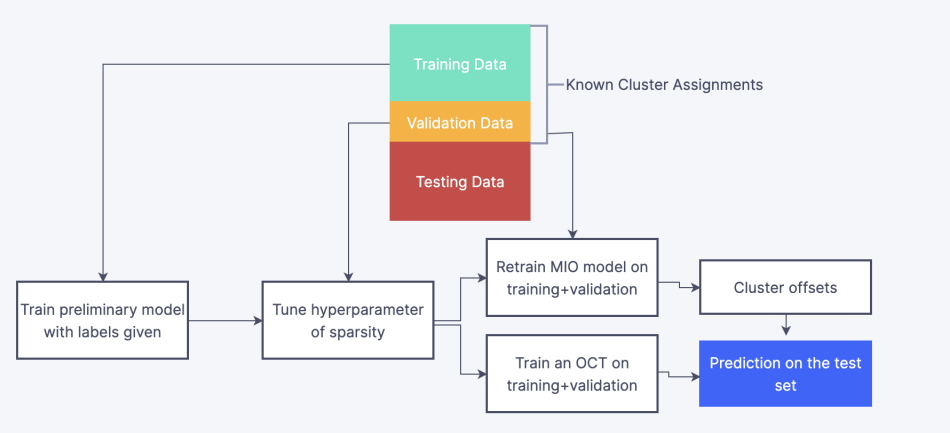

Fig 1: Pipeline for model fitting, sparsity tuning, and prediction for new clusters.

Data are partitioned into training, validation, and test sets with a cluster-blocked split so that all observations from a given cluster remain in the same fold. The MIO model is fit on the training data using known cluster assignments to estimate $(\boldsymbol{\beta}, \{\gamma_k\})$ under a candidate sparsity level. The sparsity budget(s) $\{\lambda_r\}$ are chosen by grid search using the validation set, with mean squared error as the tuning criterion (Section 2.4). After selecting $\{\lambda_r\}$, the model is refit on the union of training and validation data to obtain the final estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}_k$.

To enable prediction for previously unseen clusters, we learn a supervised mapping from cluster-level covariates to estimated effects. Specifically, using the fitted $\widehat{\boldsymbol{\gamma}}_k$ on the combined training–validation set, we train an interpretable classifier (e.g., CART or Optimal Classification Trees) to map $\mathbf{X}_k$ to a predicted cluster effect. For a new observation $\mathbf{x}_\nu$ from an

unseen cluster, we compute either a hard assignment (assign the most probable cluster and add its effect) or a soft assignment (average effects using the classifier's class probabilities). In applications, we use the soft assignment by default.

All solver settings, grid definitions for $\{\lambda_r\}$, and classifier hyperparameters are provided in the Supplement, together with pseudocode corresponding to Figure 1.

3. Simulation studies. We compare Ordinary Least Squares (OLS), Gaussian linear mixed models (LMM–Gaussian), Laplace linear mixed models (LMM–Laplace), and the proposed MIO estimator. To induce clustered structure, covariates are generated with a latent cluster effect and outcomes follow a linear model with observation noise (details in the Supplement). We consider two regimes for the random intercepts $\{\gamma_k\}$: (i) a Gaussian regime in which $\gamma_k \sim \mathcal{N}(0, \sigma_\gamma^2)$ and heterogeneity is controlled by varying σ_γ^2 across a range; and (ii) a sparse regime in which only a fraction of clusters have nonzero effects and the remaining effects are exactly zero. In the sparse regime, we vary the proportion of nonzero cluster effects over a set of levels from low to moderate sparsity; nonzero values are drawn from $\{+1, -1\}$ with equal probability. The exact range for σ_γ^2 and the sparsity levels are reported in the Supplement.

To assess scalability, we examine three data-complexity settings: *Low* (K=4, p=10, Q=0), *Medium* (K=10, p=25, Q=0), and *High* (K=14, p=35, Q=0), where K is the number of clusters, p the number of non-intercept fixed-effect covariates, and Q the number of random-effect components beyond the intercept. Each cluster has n_k=50 observations. Data are split into training, validation, and test sets with cluster-blocked partitions. Tuning for the MIO sparsity budget(s) $\{\lambda_r\}$ is performed by validation; LMMs are fit by REML; OLS serves as a baseline. We report estimation accuracy for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, selection summaries in the sparse regime, and test-set prediction on held-out clusters; additional diagnostics and the precise simulation grids appear in the Supplement.

3.1. *Estimation accuracy for fixed and random effects.* We first assess how well each method estimates the fixed effects $\boldsymbol{\beta}$ and the cluster effects $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^\top$. Accuracy is measured by the ℓ_2 estimation error, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$ and $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2$, computed on each Monte Carlo replication and summarized on a log scale. Results shown here focus on the "High" setting with $K = 14$ clusters, $p = 35$ non-intercept covariates, and $n_k = 50$ observations per cluster ($Q = 0$, random intercept only); corresponding figures for the "Low" and "Medium" settings appear in the Supplement, along with nonzero-only errors for $\boldsymbol{\gamma}$ in the sparse regime and intracluster correlation (ICC) recovery.

When the data-generating process follows the Gaussian random-intercept model (Figure 2, left panels), LMM–Gaussian performs strongly, as expected. The MIO estimator delivers comparable (and frequently lower) ℓ_2 error for $\boldsymbol{\beta}$ across heterogeneity levels, indicating that the sparsity-constrained formulation adapts well even when the Gaussian specification is correct. OLS and LMM–Laplace exhibit larger $\boldsymbol{\beta}$ errors in this regime. For $\boldsymbol{\gamma}$, the three methods that adjust for clustering (LMM–Gaussian, LMM–Laplace, and MIO) achieve similar error profiles, with small differences across settings.

Under sparse cluster effects (Figure 2, right panels), the differences are more pronounced. The MIO estimator attains noticeably lower ℓ_2 error for $\boldsymbol{\beta}$ across sparsity levels, reflecting its ability to isolate a small set of deviating clusters. OLS and LMM–Laplace show higher $\boldsymbol{\beta}$ error, and LMM–Gaussian degrades as sparsity increases. For $\boldsymbol{\gamma}$, all clustering methods perform similarly at low sparsity, while MIO yields lower error as sparsity increases. These patterns persist in the "Low" and "Medium" settings (Supplement), where we also report nonzero-only $\boldsymbol{\gamma}$ errors and selection summaries (TPR/FDP) for the sparse regime. ICC recovery results appear in the Supplement; LMM–Gaussian is near-exact under its correctly specified model, with MIO close across settings.

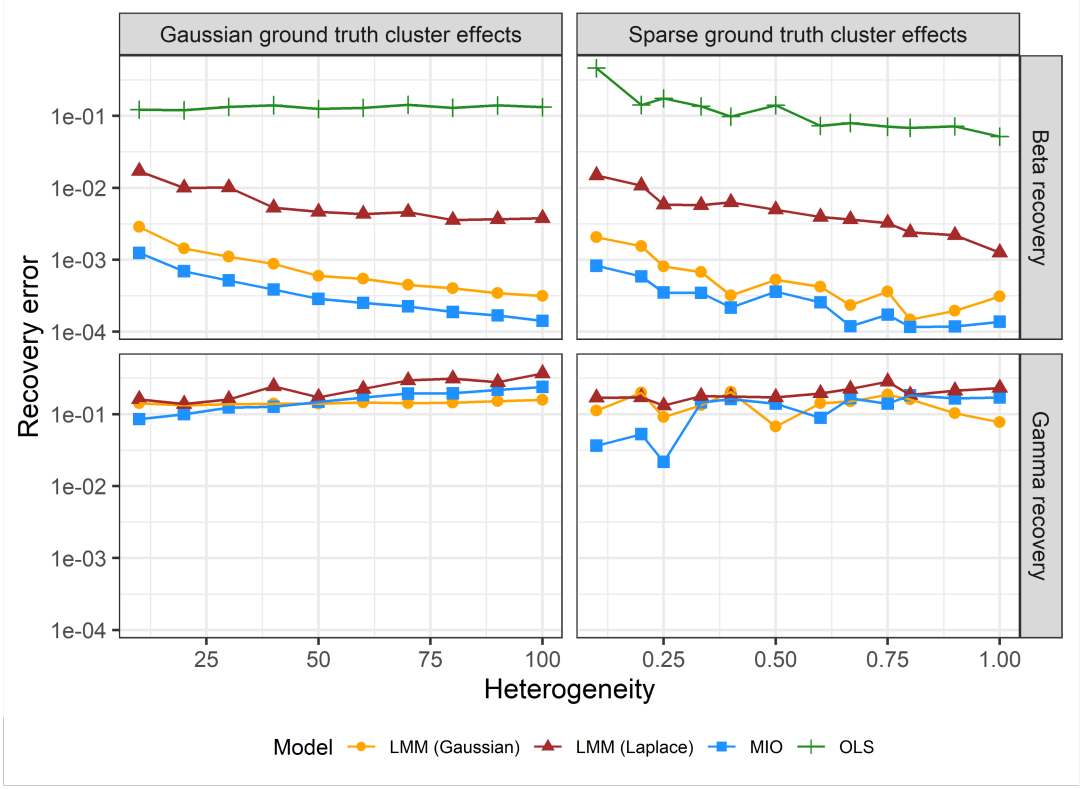

Fig 2: ℓ_2 estimation error on the log scale for fixed effects ($\boldsymbol{\beta}$, top) and cluster effects ($\boldsymbol{\gamma}$, bottom) in the high-dimensional setting (K=14, p=35, n_k=50, Q=0). Panels compare the case where cluster effects are Gaussian (left) and sparse (right).

3.2. *Predictive performance on held-out clusters.* We assess predictive accuracy on held-out clusters, corresponding to generalization to organizational units not used for model fitting. Predictions use the fitted parameters and, for methods with cluster effects, incorporate estimated or predicted $\widehat{\gamma}_k$. For MIO, soft assignment is used for new clusters (Section 2.5); LMM-based methods use empirical Bayes estimates; OLS relies on fixed effects only.

Predictive performance is summarized by test-set mean squared error (MSE) computed over clusters excluded during training. Figure 3 reports results for the high-dimensional setting (K=14, p=35, n_k=50, Q=0) under both Gaussian and sparse regimes; additional settings appear in the Supplement.

When the data-generating process follows the Gaussian random-intercept model, MIO and LMM–Gaussian yield similar test MSE, with MIO often showing modest improvements. Under the sparse regime, MIO attains lower test MSE than both LMM variants and OLS across the examined sparsity levels. Comparable patterns are observed in the "Low" and "Medium" complexity settings (Supplement). In summary, sparsity-aware selection of cluster effects tends to improve out-of-sample prediction, particularly when only a subset of clusters departs from the population mean.

3.3. *Computational cost.* We finally compare computation times across methods to evaluate their practical feasibility. All simulations were performed on a MacBook Pro (Apple M1 Pro processor) using `Gurobi 10.0.0` for the MIO solver. Average runtimes per model fit were recorded under the "Low," "Medium," and "High" complexity settings described in Section 3, with sparse data-generating processes.

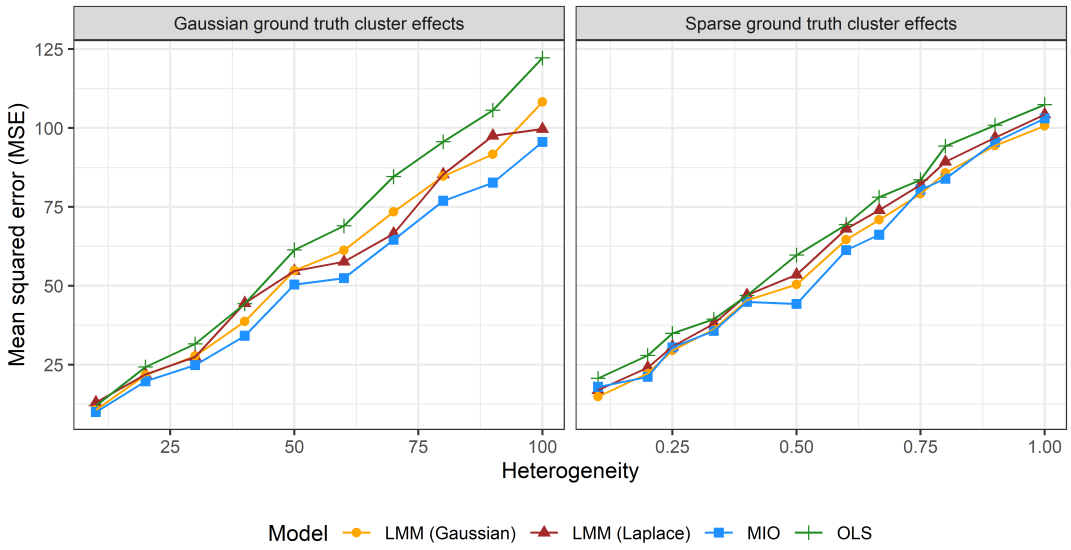

Fig 3: Test-set predictive MSE of the four algorithms in the high-dimensional setting (K=14, p=35, n_k=50, Q=0), under Gaussian (left) and sparse (right) cluster-effect regimes.

TABLE 1
Average runtime per model fit (milliseconds) across dimensionality levels under a sparse data-generating process.

Dimensionality	OLS	LMM (Gaussian)	LMM (Laplace)	MIO
Low	2.6	13.7	744.6	237.2
Medium	16.2	18.1	835.4	328.5
High	26.7	43.1	927.9	539.4

As expected, OLS and LMM–Gaussian achieve the shortest runtimes, given their closed-form or gradient-based implementations. The generalized Laplace model and MIO approach are more computationally demanding, though both remain feasible for moderate problem sizes. MIO runtimes scale roughly linearly with the number of clusters and covariates, reflecting the efficiency of modern solvers and warm-start strategies. While enforcing sparsity through integer constraints incurs additional cost, the computational burden is moderate relative to the interpretability and flexibility gains achieved by the proposed method.

4. Data examples.

4.1. *Hospital rankings.* We apply our method to the New York State Department of Health (NYSDOH) Hospital-Acquired Infections (HAI) dataset (NYSDOH, 2025), which reports annual, facility-level infection indicators for all acute-care hospitals statewide beginning in 2008. We focus on the 2019 data for hysterectomy surgical site infections (SSIs), for which hospitals report the number of infections observed, the number of procedures performed, and the number of infections predicted from statewide risk-adjustment models. Following NYSDOH methodology, we define "underperforming" hospitals as those whose infection rate is statistically higher than the New York State average after risk adjustment.

Although the public data are aggregate counts, we expand each hospital's record to pseudo–individual observations (infection $= 1$ for the observed count and 0 otherwise). Let

TABLE 2
Hospitals flagged as "significantly higher than NYS average" in 2019 hysterectomy. Comparison of MIO-selected hospitals and Gaussian LMM random intercepts.

Hospital	**State label**	**MIO effect**	**LMM (Gaussian) effect**
Faxton–St. Luke's (St. Luke's Div.)	Yes	0.0459	0.0113
Jacobi Medical Center	Yes	0.0460	0.0092
Montefiore—Jack D. Weiler	Yes	0.0339	0.0153
Next-highest (e.g., Oswego)	No	—	0.0069
Next-highest (e.g., N. Westchester)	No	—	0.0065

ω_k denote the NYSDOH risk–adjusted expected infection rate for hospital k. We then fit a LMM on the probability scale,

$$Y_{ki} = \mu + \omega_k + \gamma_k + \varepsilon_{ki}, \qquad \varepsilon_{ki} \sim \mathcal{N}(0, \sigma^2), \quad \gamma_k \sim F_\gamma,$$

so that γ_k captures the hospital–specific deviation from the statewide risk–adjusted benchmark. This aligns with our MIO formulation in Section 2.3 (random intercepts, $Q{=}0$) and yields effects interpretable as probability differences. While $Y_{ki} \in \{0, 1\}$ is binary, using a LMM in place of a logistic GLMM is often acceptable—particularly with few covariates, large sample sizes, and probabilities away from the boundaries, and provides unbiased estimates of average marginal effects under broad conditions (see, e.g., Chen et al., 2023; Gomila, 2021).

Figure 4 displays the estimated random intercepts from the Gaussian LMM fit, sorted across hospitals. Three hospitals stand out with substantially positive effects—Faxton–St. Luke's Healthcare (St. Luke's Division), Jacobi Medical Center, and Montefiore Medical Center (Jack D. Weiler Hospital) matching those labeled by the NYSDOH as "significantly higher than the state average." We set the MIO sparsity level $\lambda_0 = 3$ to correspond to the number of state-flagged hospitals and refit the model on the selected support to remove shrinkage bias. Table 2 compares the hospitals identified by MIO to those selected by the Gaussian LMM.

The MIO selection matches the NYSDOH public classification, while the effect magnitudes differ from Gaussian LMM estimates for a principled reason. A Gaussian LMM imposes an ℓ_2 penalty which yields continuous shrinkage toward zero, which attenuates, but never exactly sets, random effects γ_k to zero. In contrast, our SCS formulation enforces an ℓ_0 sparsity constraint $\|\gamma'\|_0 \leq \lambda_0$, producing hard selection at the hospital level (many exact zeros) and estimating the retained γ_k without penalty–induced shrinkage; selected effects are therefore typically larger (less attenuated). The optimizer's binary inclusion/exclusion can be viewed as a one–step selection decision at the chosen λ_0 rather than a classical confidence interval. If interval estimates are desired, they can be added post hoc (e.g., cluster–level bootstrap on the fitted solution), but the primary output here is the discrete selection and the associated effect estimates for selected hospitals.

4.2. *Student performance.* We analyze the Portuguese secondary school dataset of Cortez and Silva (2008), comprising $N = 382$ students with demographic, social, and school-related predictors (e.g., parental education, study time). Continuous variables were standardized and categorical variables encoded with dummies, yielding $p = 20$ non-intercept covariates. We fit separate models for Mathematics and Portuguese test scores. Students are treated as clusters with a random intercept ($Q{=}0$), so that the student-specific effect γ_k captures adjusted performance after accounting for observed covariates.

Our goal is to identify students whose adjusted performance deviates meaningfully from the population mean. For comparison with regression-free ranking, we note that ordering by

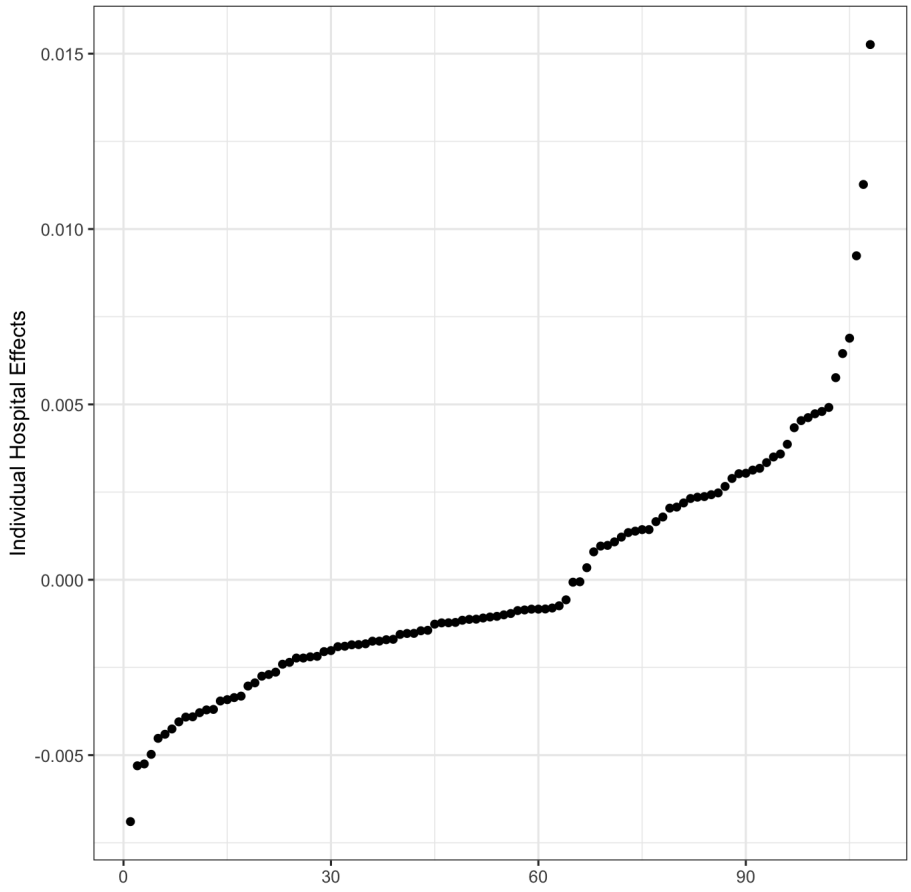

Fig 4: Sorted hospital random intercepts ($\hat{\gamma}_k$) from a Gaussian LMM for 2019 hysterectomy infections. Hospitals with higher positive effects indicate increased infection risk relative to the statewide benchmark.

raw average scores can be misleading because it ignores differences in background and school environment. All methods use the same fixed-effects specification and a cluster-blocked split at the student level (train/validation/test). The sparsity level λ_0 for MIO is chosen by validation (Section 2.4); LMMs are fit by REML; OLS is included as a baseline.

Figure 5 (left panel) displays the estimated random intercepts for each student under the Laplace LMM, Gaussian LMM, and MIO. The MIO estimates exhibit many exact zeros with a subset of clearly nonzero effects, reflecting the ℓ_0 selection constraint that sets unselected students' effects to zero and leaves selected effects less attenuated. In contrast, Gaussian and Laplace LMMs produce continuously shrunk estimates (BLUPs) due to their implicit ℓ_2-type regularization, yielding values closer to zero and rarely exactly zero.

Figure 5 (right panel) compares adjusted student effects to raw average scores via QQ-style summaries. While the Gaussian and Laplace LMM effects track raw scores closely, MIO often reorders the tails: some students with extreme raw scores are assigned $\widehat{\gamma}_k = 0$ after adjustment, whereas some with middling raw scores are flagged as having nonzero adjusted effects. This difference is consistent with the roles of shrinkage versus selection: BLUPs continuously pull all effects toward zero, whereas the SCS formulation makes a discrete inclusion decision at the chosen λ_0. Overlap of top/bottom deciles across methods and

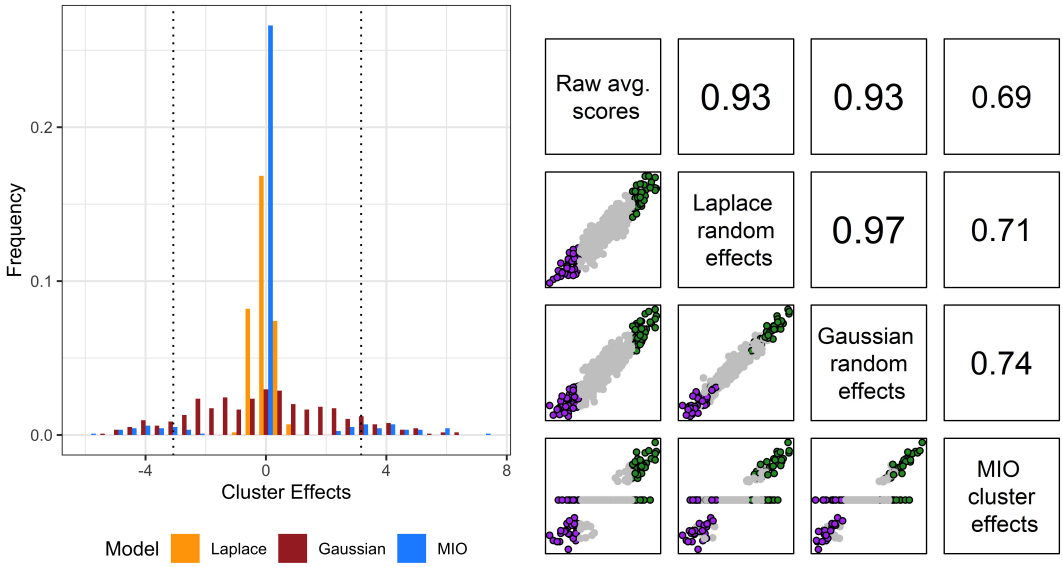

Fig 5: (Left) Predicted random effects assuming Laplace-distributed, Gaussian-distributed, and MIO and (Right) Concordance of cluster effects with raw scores across different algorithms. The points are colored based on the raw average scores (green = top 10%, purple = bottom 10%)

rank correlations are reported in the Supplement, alongside test-set RMSE for out-of-sample prediction.

4.3. *Protein expression.* We also consider a mouse protein expression dataset (Dua and Graff, 2017; Ahmed et al., 2015) and the analysis framework of Higuera et al. (2015) as a brief supplementary example. Observations are grouped into clusters (e.g., animal or experimental condition), and we compare predictive performance across OLS, LMMs, and MIO using a common fixed-effects design with cluster-blocked evaluation. In this setting, OLS does not leverage cluster structure and performs worse out of sample; a Gaussian LMM encounters convergence issues in several splits (near-singular fits), whereas the MIO formulation remains stable and achieves lower test error by selecting a sparse set of cluster effects. Full details—including data preprocessing, model specifications, convergence diagnostics, and test RMSE summaries—are provided in the Supplement.

5. Discussion. We studied significant cluster selection (SCS) in linear mixed models via a mixed-integer formulation that enforces ℓ_0 sparsity on random effects. Across Gaussian and sparse regimes in simulations, SCS identified atypical clusters while maintaining competitive fixed–effect estimation and strong predictive performance on held-out clusters. Two data illustrations—hospital-acquired infection reporting and student performance—showed how SCS yields discrete, interpretable selections that align with external benchmarks or reveal adjusted differences not visible from raw rankings.

In settings with approximately Gaussian random effects, SCS performed on par with Gaussian LMM for fixed–effect estimation and prediction; in sparse regimes, SCS reduced estimation error and improved prediction by concentrating mass on a small set of clusters. For random-effect recovery, SCS was most advantageous when true heterogeneity was sparse

or modest in variance; when heterogeneity was pervasive with high variance, continuous-shrinkage LMMs were competitive and selection became more sensitive to the sparsity budget. We examined stability across nearby choices of λ_r and report certified solver gaps and compute times in the Supplement.

A central distinction between SCS and Gaussian LMM concerns shrinkage versus selection. Gaussian LMMs report BLUPs, which are algebraically equivalent to ℓ_2 penalization on random effects and therefore exhibit continuous shrinkage toward zero; estimates are attenuated and rarely exactly zero. SCS imposes an ℓ_0 constraint, yielding exact zeros for non-selected clusters and less attenuation for those retained. Consequently, SCS returns a sparse active set with larger effect magnitudes for selected clusters, whereas BLUPs produce smaller, nonzero estimates for nearly all clusters. In applications where the goal is to flag a limited number of atypical units for review, this discrete selection can be especially useful.

When outcomes are binary but covariate structure is limited and probabilities are moderate, we demonstrate a linear mixed model on the probability scale can be adequate. In institutional ranking problems (e.g., hospitals), SCS selections should be interpreted as a screening tool to prioritize review rather than as definitive determinations.

This work has limitations and natural extensions. Our analyses use known cluster labels and random intercepts; classroom/teacher effects in the student data and random slopes in the hospital context were not modeled. Selections depend on λ_r and may vary near boundary settings, although we observe stability across practical ranges. Methodologically, grouped budgets across random-effect components, GLM losses with the same ℓ_0 selection mechanism, and multivariate outcomes with shared cluster effects are all feasible within the MIO framework but beyond the present scope. Future work includes post-selection uncertainty quantification for random effects (e.g., bootstrap intervals on the selected support), procedures for false discovery rate control over selected clusters, and handling unknown cluster assignments by combining probabilistic assignment with SCS.

Software and data availability. Julia code implementing SCS (model fitting, tuning, and prediction), simulation scripts, and replication materials are available at: https://github.com/Madhav1812/cluster-mio. The New York State Department of Health Hospital-Acquired Infections (HAI) dataset used in the hospital illustration is publicly available via Health Data NY.

REFERENCES

Ahmed MM, Dhanasekaran AR, Block A, Tong S, Costa ACS, Stasko M, Gardiner KJ (2015) Protein dynamics associated with failed and rescued learning in the ts65dn mouse model of down syndrome. *PLOS ONE* 10:e0119491.

Bertsimas D, Dunn J (2017) Optimal classification trees. *Machine Learning* 106:1039–1082.

Bertsimas D, King A, Mazumder R (2015) Best subset selection via a modern optimization lens arXiv:1507.03133 [math, stat].

Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) *Classification and regression trees* Routledge.

Chen K, Martin RS, Wooldridge JM (2023) Another look at the linear probability model and nonlinear index models. *arXiv preprint arXiv:2308.15338* .

Chen T, Tchetgen Tchetgen EJ, Wang R (2020) A stochastic second-order generalized estimating equations approach for estimating association parameters. *Journal of Computational and Graphical Statistics* 29:547–561.

Chen T, Wang R (2020) Inference for variance components in linear mixed-effect models with flexible random effect and error distributions. *Statistical Methods in Medical Research* 29:3586–3604.

Chetty R, Friedman JN, Rockoff JE (2014) Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review* 104:2633–2679.

Cortez P, Silva AMG (2008) Using data mining to predict secondary school student performance .

Del Pia A, Weismantel R (2010) On convergence in mixed integer programming. *Mathematical Programming* 135:1–16.

Dua D, Graff C (2017) UCI machine learning repository.

Duran MA, Grossmann IE (1986) An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical programming* 36:307–339.

Fukushima M (1984) On the convergence of a class of outer approximation algorithms for convex programs. *Journal of Computational and Applied Mathematics* 10:147–156.

Geraci M, Farcomeni A (2020) A family of linear mixed-effects models using the generalized laplace distribution. *Statistical Methods in Medical Research* 29:2665–2682.

Gomila R (2021) Logistic or linear? estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General* 150:700.

Greenshtein E (2006) Best subset selection, persistence in high-dimensional statistical learning and optimization under l 1 constraint .

Gurobi Optimization, LLC (2023) Gurobi Optimizer Reference Manual.

Higuera C, Gardiner KJ, Cios KJ (2015) Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PLOS ONE* 10:e0129126.

Hota B, Webb T, Chatrathi A, McAninch E, Lateef O (2020) Disagreement between hospital rating systems: measuring the correlation of multiple benchmarks and developing a quality composite rank. *American Journal of Medical Quality* 35:222–230.

Kartal Koc E, Bozdogan H (2015) Model selection in multivariate adaptive regression splines (mars) using information complexity as the fitness function. *Machine Learning* 101:35–58.

Kowal DR (2023) Subset selection for linear mixed models. *Biometrics* 79:1853–1867.

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. *Biometrics* pp. 963–974.

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22.

Lin TI (2008) Longitudinal data analysis using t linear mixed models with autoregressive dependence structures. *Journal of Data Science* 6:333–355.

Lin TI, Lee JC (2008) Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. *Statistics in medicine* 27:1490–1507.

Mazumder R, Friedman JH, Hastie T (2011) Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association* 106:1125–1138.

McCaffrey DF, Lockwood JR, Koretz D, Louis TA, Hamilton L (2004) Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics* 29:67–101.

Ntani G, Inskip H, Osmond C, Coggon D (2021) Consequences of ignoring clustering in linear regression. *BMC Medical Research Methodology* 21:139.

NYSDOH (2025) Hospital-acquired infections (hai) — hospital-level data Facility-level HAI indicators; we use the 2019 hysterectomy SSI subset.

Oberauer K (2022) The importance of random slopes in mixed models for bayesian hypothesis testing. *Psychological Science* 33:648–665.

Pinheiro JC, Liu C, Wu YN (2001) Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* 10:249–276.

Raskutti G, Wainwright MJ, Yu B (2011) Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory* 57:6976–6994.

Verbeke G, Lesaffre E (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* 91:217–221.

Zhang Y, Wainwright MJ, Jordan MI (2014) Lower bounds on the performance of polynomial-time algorithms for sparse linear regression In *Conference on Learning Theory*, pp. 921–948. PMLR.

SUPPLEMENT TO "A DISTRIBUTION-FREE MIXED-INTEGER OPTIMIZATION APPROACH TO HIERARCHICAL MODELLING OF CLUSTERED AND LONGITUDINAL DATA"

A. Theoretical results. The convergence of our algorithm can be proven using results from previous works on best subset selection. For ease of interpretation, the following results will hold for the random intercepts model described in Section 2.1. The results are generalizable and can be extended for a generic number of random effects as well.

DEFINITION 1. If $\boldsymbol{B}$ is a matrix, we define $\boldsymbol{B}^\intercal\boldsymbol{B}$ to be the corresponding Gram matrix.

DEFINITION 2. The spectrum of a square matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ is the set of eigenvalues of $\{\lambda_i(\boldsymbol{B})\}_{i=1}^n$, where $\lambda_1 \geq \cdots \geq \lambda_n$. By definition, we always take $\lambda_1(\mathbf{B})$ to be the largest eigenvalue of $\mathbf{B}$.

In order to tackle the main convergence results, we must establish some details regarding the properties of the design matrix, specifically, the spectrum of its corresponding Gram matrix.

PROPOSITION 1. Suppose $\lambda_1(\mathbf{X}^\intercal\mathbf{X}) = \mathcal{O}(n)$, where $n = \sum_{k=1}^K n_k$. Then $\lambda_1(\widetilde{\boldsymbol{X}}^\intercal\widetilde{\boldsymbol{X}}) = \mathcal{O}(n)$.

PROOF. First, note that $\mathbf{A}^\intercal\mathbf{A} = \text{diag}(n_1, \cdots, n_K)$, since $\mathbf{A}$ mimics an incidence matrix with individuals representing vertices and clusters representing edges in a hypergraph, and therefore $\mathbf{A}^\intercal\mathbf{A}$ enumerates the size of each cluster. Next, let

$$\mathbf{N} = \begin{bmatrix} \boldsymbol{X}^\intercal\boldsymbol{X} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}^\intercal\boldsymbol{A} \end{bmatrix} \quad \text{and} \quad \mathbf{R} = \begin{bmatrix} \mathbf{0} & \boldsymbol{X}^\intercal\boldsymbol{A} \\ \boldsymbol{A}^\intercal\boldsymbol{X} & \mathbf{0} \end{bmatrix}$$

Then by Weyl's inequality,

$$\lambda_1(\widetilde{\boldsymbol{X}}^\intercal\widetilde{\boldsymbol{X}}) \leq \lambda_1(\mathbf{N}) + \lambda_1(\mathbf{R})$$

Due to the block-diagonal structure of $\mathbf{N}$, it immediately follows that the spectrum of $\mathbf{N}$ is the union of the spectrums of $\mathbf{X}^\intercal\mathbf{X}$ and $\mathbf{A}^\intercal\mathbf{A}$, and therefore $\lambda_1(\mathbf{N}) = \max\{\mathcal{O}(n), \max\{n_1, \cdots, n_K\}\} = \mathcal{O}(n)$.

Note that the eigenvalues of $\mathbf{R}$ are precisely the singular values of $\mathbf{A}^\intercal\mathbf{X}$ (with additional multiplicity); denote $\sigma_i(\mathbf{A}^\intercal\mathbf{X})$ as these singular values, also in decreasing order $\sigma_1 \geq \cdots \geq \sigma_{\min\{K,P+1\}}$. We then have

$$\sigma_1(\mathbf{A}^\intercal\mathbf{X}) \leq \sigma_1(\mathbf{A})\sigma_1(\mathbf{X}) = \sqrt{\lambda_1(\mathbf{A}^\intercal\mathbf{A})\lambda_1(\mathbf{X}^\intercal\mathbf{X})} = \sqrt{\max\{n_1, \cdots, n_K\} \cdot \mathcal{O}(n)} = \mathcal{O}(n)$$

□

In order to introduce the following theorem, we need the following setup for the optimization problem. Let us assume that the optimization problem is of the form:

$$\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}) \text{ subject to } ||\boldsymbol{\beta}||_0 \leq k$$

Keywords and phrases: clustered data, significant cluster selection, mixed integer programming, random effect ranking.

where g is a convex function, and has a ℓ-Lipschitz gradient, that is

$$||\nabla g(\boldsymbol{\beta}) - \nabla g(\widetilde{\boldsymbol{\beta}})|| \leq \ell ||\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}||$$

In the case of the least squares setup we have, the explicit form of ℓ is given by $\lambda_1(\boldsymbol{X}^\intercal \boldsymbol{X})$, where X is the design matrix. Thus, from the previous proposition, the Lipschitz constant grows at the same rate as a normal regression of Y on $\boldsymbol{X}$.

Now, the approach to solving this problem for a general convex function g is to use an upper bound of simpler form for the function, and minimize over this upper bound. This is the essence of the outer approximation algorithm. The construction of these upper bounds requires a variable $L > \ell$ which represents the coarseness of the approximation, and in term, the "step size" of the descent algorithm (larger L, finer approximation).

Thus, under the algorithm given in Bertsimas et al. (2015), the following theorem follows.

THEOREM 1 (Theorem 3.1, Bertsimas et al. (2015)). Let $L > \ell$ and $\boldsymbol{\beta}^*$ be the optimum of the optimization function. After M iterations, we have

$$\min_{m=1,\cdots,M} \|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m\|_2^2 \leq \frac{2(g(\boldsymbol{\beta}) - g(\boldsymbol{\beta}^*))}{M(L-\ell)}$$

The convergence rate of our algorithm is contingent on the complexity of the function g and the magnitude of ℓ. According to our proposition, we have adjusted the growth rate of ℓ to align with the rate provided in the associated paper, with g representing the straightforward least squares error function.

The underlying principle of Cauchy convergence implies a trajectory towards the optimum solution, given that we consider a restricted complete support for our vector. This results in a framework of convergence for our algorithm towards a global optimum.

Furthermore, we wish to highlight the efficiency of the algorithm in retrieving non-zero coefficients, also referred to as the "active set". The algorithm, by design, initializes the first p coordinates as non-sparse, along with λ proportion of the remaining coordinates. This strategy enables the algorithm to utilize outer approximation cuts to accurately identify the real active set, followed by a phase of refining the effect sizes. Empirical observations corroborated this phenomenon, providing a rationale for limiting the algorithm's iterations as a means to curtail computation time.

B. Simulation results. We provide further simulation results that were not included in the main body of the paper. All results are presented in tabular form as well. The results corroborate the conclusions we drew in the main body of the paper. In particular, compared to the other approaches, the MIO approach is generally more effective in the recovery of the causal parameters of interest ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$), and can subsequently leverage the parameters for improved prediction (i.e. better MSE on the test set), and inference (i.e. correct recovery of sparsity/ICC).

B.1. *Cluster effects are truly sparse.*

B.1.1. *β and γ recovery.*

B.1.2. *Sparsity recovery.* It should be noted, that the competing methods (Laplace LMM and Gaussian LMM) cannot recovery sparse cluster effects γ_k, i.e. inferred sparsity of γ_k is always 0%, regardless of what the true sparsity is. In comparison, we see efficient recovery of the sparsity my the MIO approach, and this recovery even improves with the dimensionality of the regresion problem.

TABLE B1

Beta recovery under the simulation scenarios where the cluster-effects are truly sparse, with low dimensionality (4 clusters, 10 covariates, 50 observations per cluster)

True sparsity in γ_k	$\|\|\beta_{\text{true}} - \widehat{\beta}_{OLS}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(G)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(L)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{MIO}\|\|_2^2$
90%	0.23564	0.00122	0.01664	**0.00054**
80%	0.23334	0.00116	0.01381	**0.00055**
75%	0.23536	0.00121	0.01248	**0.00052**
66%	0.10705	0.00056	0.00981	**0.00024**
60%	0.10894	0.00060	0.00781	**0.00028**
50%	0.10477	0.00061	0.00546	**0.00029**
40%	0.07245	0.00038	0.00310	**0.00019**
33%	0.07114	0.00036	0.00167	**0.00018**
25%	0.07129	0.00035	0.00098	**0.00018**
20%	0.04846	0.00025	0.00083	**0.00013**
10%	0.05316	0.00030	0.00071	**0.00014**
0%	0.05406	0.00024	0.00052	**0.00013**

TABLE B2

Beta recovery under the simulation scenarios where the cluster-effects are truly sparse, with medium dimensionality (10 clusters, 25 covariates, 50 observations per cluster)

True sparsity in γ_k	$\|\|\beta_{\text{true}} - \widehat{\beta}_{OLS}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(G)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(L)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{MIO}\|\|_2^2$
90%	0.29369	0.00402	0.01492	**0.00143**
80%	0.28198	0.00159	0.00877	**0.00069**
75%	0.18567	0.00099	0.00462	**0.00045**
66%	0.13559	0.00073	0.00436	**0.00034**
60%	0.14095	0.00071	0.00419	**0.00035**
50%	0.11050	0.00056	0.00477	**0.00026**
40%	0.09078	0.00045	0.00467	**0.00022**
33%	0.07682	0.00038	0.00342	**0.00018**
25%	0.06779	0.00034	0.00326	**0.00017**
20%	0.06986	0.00034	0.00338	**0.00016**
10%	0.05724	0.00029	0.00317	**0.00014**
0%	0.05325	0.00027	0.00296	**0.00014**

TABLE B3

Beta recovery under the simulation scenarios where the cluster-effects are truly sparse, with high dimensionality (14 clusters, 35 covariates, 50 observations per cluster)

True sparsity in γ_k	$\|\|\beta_{\text{true}} - \widehat{\beta}_{OLS}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(G)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(L)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{MIO}\|\|_2^2$
90%	0.45927	0.00205	0.01489	**0.00082**
80%	0.14163	0.00154	0.01074	**0.00059**
75%	0.17521	0.00080	0.00583	**0.00035**
66%	0.13538	0.00067	0.00576	**0.00034**
60%	0.09782	0.00032	0.00629	**0.00022**
50%	0.14049	0.00053	0.00497	**0.00036**
40%	0.07279	0.00042	0.00394	**0.00025**
33%	0.07975	0.00023	0.00364	**0.00012**
25%	0.07071	0.00036	0.00324	**0.00017**
20%	0.06822	0.00015	0.00240	**0.00012**
10%	0.07138	0.00019	0.00219	**0.00012**
0%	0.05130	0.00031	0.00126	**0.00014**

B.2. *Cluster effects are truly Gaussian.*

B.2.1. *β and γ recovery.*

TABLE B4

Gamma recovery for cluster-based methods under the simulation scenarios where the cluster-effects are truly sparse, with low dimensionality (4 clusters, 10 covariates, 50 observations per cluster)

True sparsity in γ_k	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(G)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(L)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{MIO}\|\|_2^2$
90%	0.03988	0.05686	**0.02347**
80%	0.04573	0.07390	**0.02194**
75%	0.04849	0.06127	**0.02554**
66%	0.04519	0.11464	**0.03990**
60%	0.04687	0.09935	**0.03587**
50%	0.04173	0.11481	**0.04027**
40%	0.05245	0.19631	**0.04295**
33%	0.04680	0.08649	**0.04345**
25%	0.05033	0.09610	**0.04175**
20%	**0.04401**	0.09123	0.05061
10%	**0.03825**	0.06293	0.05017
0%	**0.04061**	0.11586	0.05236

TABLE B5

Gamma recovery for cluster-based methods under the simulation scenarios where the cluster-effects are truly sparse, with medium dimensionality (10 clusters, 25 covariates, 50 observations per cluster)

True sparsity in γ_k	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(G)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(L)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{MIO}\|\|_2^2$
90%	0.11046	0.08674	**0.03191**
80%	0.09521	0.07189	**0.05322**
75%	0.10700	0.14838	**0.06776**
66%	0.10952	0.15673	**0.07681**
60%	0.10673	0.14789	**0.07519**
50%	0.10657	0.16863	**0.08277**
40%	0.11001	0.18258	**0.10667**
33%	0.10921	0.16142	**0.10074**
25%	0.10388	0.17106	**0.09744**
20%	**0.10748**	0.15626	0.12749
10%	**0.10500**	0.17360	0.14658
0%	**0.12681**	0.16553	0.14561

TABLE B6

Gamma recovery for cluster-based methods under the simulation scenarios where the cluster-effects are truly sparse, with high dimensionality (14 clusters, 35 covariates, 50 observations per cluster)

True sparsity in γ_k	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(G)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(L)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{MIO}\|\|_2^2$
90%	0.11147	0.16897	**0.03638**
80%	0.19905	0.17153	**0.05270**
75%	0.09119	0.13122	**0.02171**
66%	**0.13254**	0.17776	0.14525
60%	0.20280	0.17549	**0.16220**
50%	**0.06703**	0.17158	0.13927
40%	0.14134	0.19434	**0.00887**
33%	**0.15031**	0.22395	0.16627
25%	0.18617	0.28267	**0.13943**
20%	**0.15981**	0.18452	0.18249
10%	**0.10285**	0.21238	0.16508
0%	**0.07705**	0.23086	0.16996

B.2.2. *Intra-cluster correlation (ICC) recovery.* We analyze the Intraclass Correlation Coefficient (ICC) recovery of our algorithm in comparison to the baseline methodologies. ICC serves as a crucial parameter given its capacity to capture intra-cluster observation be-

TABLE B7
Sparsity recovery of Gamma of the MIO approach under the different dimensional setups

True sparsity in γ_k	Low	Medium	High
90%	79%	82%	86%
80%	70%	71%	79%
75%	60%	65%	72%
66%	57%	59%	63%
60%	49%	53%	58%
50%	41%	43%	45%
40%	38%	33%	39%
33%	25%	25%	31%
25%	15%	15%	20%
20%	11%	14%	15%
10%	0%	6%	8%
0%	0%	0%	0%

TABLE B8
Beta recovery under the simulation scenarios where the cluster-effects are truly Gaussian, with low dimensionality (4 clusters, 10 covariates, 50 observations per cluster)

Variance of γ_k	$\|\|\beta_{\text{true}} - \widehat{\beta}_{OLS}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(G)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(L)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{MIO}\|\|_2^2$
10.0	0.09206	0.00378	0.04470	**0.00175**
20.0	0.10171	0.00199	0.03591	**0.00089**
30.0	0.10263	0.00132	0.02300	**0.00072**
40.0	0.10711	0.00134	0.01236	**0.00061**
50.0	0.09984	0.00094	0.01030	**0.00047**
60.0	0.10761	0.00111	0.00796	**0.00052**
70.0	0.11339	0.00060	0.00682	**0.00027**
80.0	0.10867	0.00069	0.00378	**0.00031**
90.0	0.10310	0.00049	0.00154	**0.00026**
100.0	0.10034	0.00045	0.00078	**0.00020**

TABLE B9
Beta recovery under the simulation scenarios where the cluster-effects are truly Gaussian, with medium dimensionality (10 clusters, 25 covariates, 50 observations per cluster)

Variance of γ_k	$\|\|\beta_{\text{true}} - \widehat{\beta}_{OLS}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(G)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(L)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{MIO}\|\|_2^2$
10.0	0.11150	0.00289	0.00947	**0.00142**
20.0	0.12713	0.00161	0.00656	**0.00074**
30.0	0.11915	0.00116	0.00569	**0.00056**
40.0	0.13113	0.00094	0.00468	**0.00043**
50.0	0.12789	0.00068	0.00477	**0.00034**
60.0	0.11203	0.00061	0.00463	**0.00029**
70.0	0.12111	0.00045	0.00475	**0.00020**
80.0	0.13204	0.00042	0.00409	**0.00020**
90.0	0.12829	0.00042	0.00376	**0.00020**
100.0	0.13437	0.00035	0.00429	**0.00016**

havior, which has broad applications in various clinical settings, such as assessing hospital homogeneity.

Our simulations illustrate that the Linear Mixed Model (LMM) with Gaussian cluster effects accurately recovers the true ICC, which is anticipated given its alignment with the underlying data generation process. Intriguingly, our Mixed Integer Optimization (MIO) model also demonstrates commendable ICC recovery, especially when contrasted with the LMM fitted with Laplace cluster effects. The MIO exhibits enhanced recovery for higher ICC lev-

TABLE B10

Beta recovery under the simulation scenarios where the cluster-effects are truly Gaussian, with high dimensionality (14 clusters, 35 covariates, 50 observations per cluster)

Variance of γ_k	$\|\|\beta_{\text{true}} - \widehat{\beta}_{OLS}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(G)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{LMM(L)}\|\|_2^2$	$\|\|\beta_{\text{true}} - \widehat{\beta}_{MIO}\|\|_2^2$
10.0	0.12196	0.00285	0.01704	**0.00124**
20.0	0.11948	0.00143	0.01000	**0.00069**
30.0	0.13324	0.00110	0.01011	**0.00052**
40.0	0.14063	0.00087	0.00529	**0.00039**
50.0	0.12514	0.00060	0.00463	**0.00029**
60.0	0.12936	0.00054	0.00432	**0.00025**
70.0	0.14265	0.00045	0.00459	**0.00023**
80.0	0.12956	0.00040	0.00356	**0.00019**
90.0	0.13977	0.00034	0.00364	**0.00017**
100.0	0.13182	0.00031	0.00378	**0.00014**

TABLE B11

Gamma recovery for cluster-based methods under the simulation scenarios where the cluster-effects are truly Gaussian, with low dimensionality (4 clusters, 10 covariates, 50 observations per cluster)

Variance of γ_k	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(G)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(L)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{MIO}\|\|_2^2$
10.0	0.04014	0.07613	**0.02846**
20.0	0.04251	0.13257	**0.02961**
30.0	0.04119	0.10150	**0.03812**
40.0	0.04579	0.14589	**0.04118**
50.0	**0.03857**	0.17664	0.04495
60.0	**0.04116**	0.08313	0.04400
70.0	**0.04085**	0.08387	0.05030
80.0	**0.03746**	0.17306	0.05121
90.0	**0.04510**	0.14564	0.06128
100.0	**0.04305**	0.14890	0.06324

TABLE B12

Gamma recovery for cluster-based methods under the simulation scenarios where the cluster-effects are truly Gaussian, with medium dimensionality (10 clusters, 25 covariates, 50 observations per cluster)

Variance of γ_k	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(G)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(L)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{MIO}\|\|_2^2$
10.0	0.10519	0.17888	**0.07696**
20.0	0.09958	0.15849	**0.09785**
30.0	0.10009	0.13542	**0.08733**
40.0	0.10931	0.13914	**0.10110**
50.0	**0.10460**	0.17986	0.11122
60.0	**0.09965**	0.16779	0.11120
70.0	**0.10389**	0.15837	0.13616
80.0	**0.10631**	0.17409	0.14843
90.0	**0.10476**	0.18811	0.15833
100.0	**0.10682**	0.18989	0.16284

els, signifying stronger within-group interactions. This marked improvement in the MIO's performance underscores its effectiveness in handling more complex interaction dynamics.

B.3. *Further simulation plots.* Fig B1-B4 demonstrate other plots from our simulation studies.

TABLE B13

Gamma recovery for cluster-based methods under the simulation scenarios where the cluster-effects are truly Gaussian, with high dimensionality (14 clusters, 35 covariates, 50 observations per cluster)

Variance of γ_k	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(G)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{LMM(L)}\|\|_2^2$	$\|\|\gamma_{\text{true}} - \widehat{\gamma}_{MIO}\|\|_2^2$
10.0	0.14146	0.16100	**0.08520**
20.0	0.13318	0.13843	**0.09959**
30.0	0.13711	0.16041	**0.12369**
40.0	0.14007	0.24408	**0.12649**
50.0	**0.13960**	0.17184	0.14740
60.0	**0.14571**	0.22432	0.16954
70.0	**0.14166**	0.29360	0.19313
80.0	**0.14382**	0.31070	0.19481
90.0	**0.15180**	0.27851	0.21748
100.0	**0.15772**	0.36796	0.23885

TABLE B14

Comparison of ICC recovery in Medium dimensionality (10 clusters, 25 covariates, 50 observations per cluster) between LMM (Gaussian) and MIO

True ICC	LMM (Gaussian) ICC	LMM (Laplace) ICC	MIO ICC
10%	8%	1%	4%
20%	18%	8%	13%
30%	29%	20%	25%
40%	39%	31%	39%
50%	49%	41%	48%
60%	59%	50%	59%
70%	70%	59%	69%
80%	80%	68%	79%
90%	90%	89%	90%

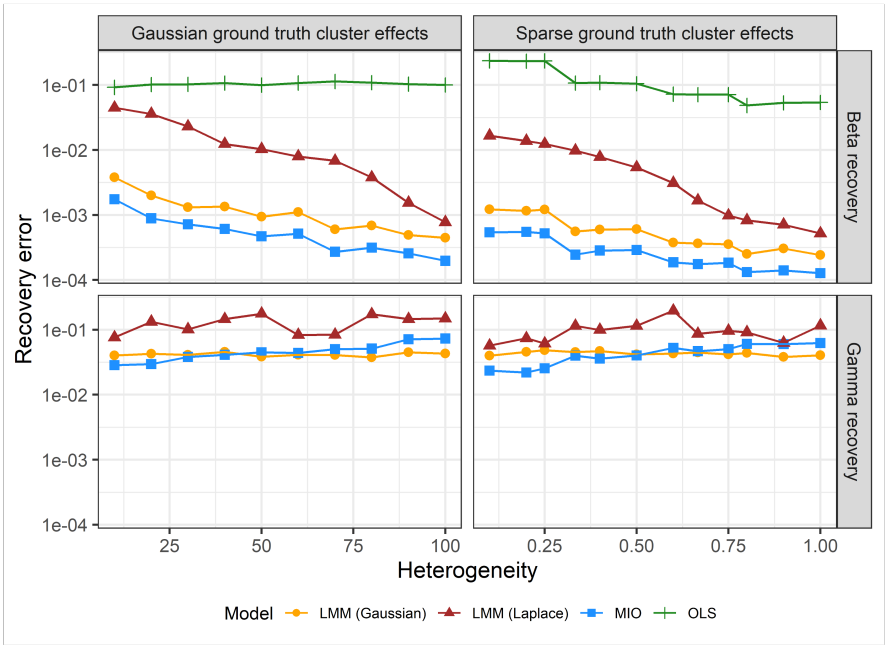

Fig B1: ℓ_2 causal effect recovery on the log scale (β (above) and γ (below)) under the simulation scenarios where the cluster-effects are truly Gaussian (left) and truly sparse (right), with low dimensional setup (4 clusters, 10 covariates, 50 observations per cluster), where the cluster-effects are truly Gaussian (left) and truly sparse (right)

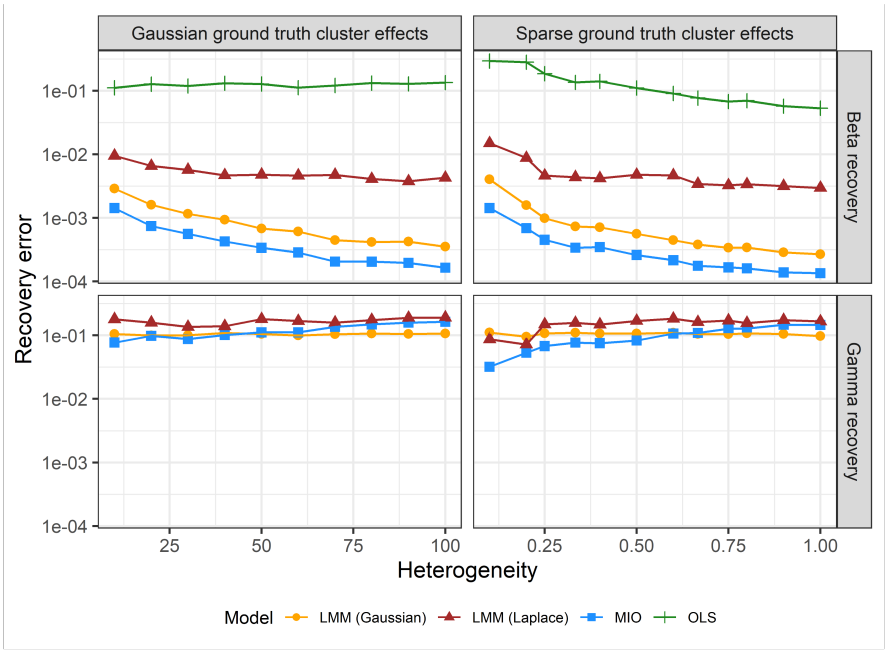

Fig B2: ℓ_2 causal effect recovery on the log scale (β (above) and γ (below)) under the simulation scenarios where the cluster-effects are truly Gaussian (left) and truly sparse (right), with a medium dimensional setup (10 clusters, 25 covariates, 50 observations per cluster), where the cluster-effects are truly Gaussian (left) and truly sparse (right)

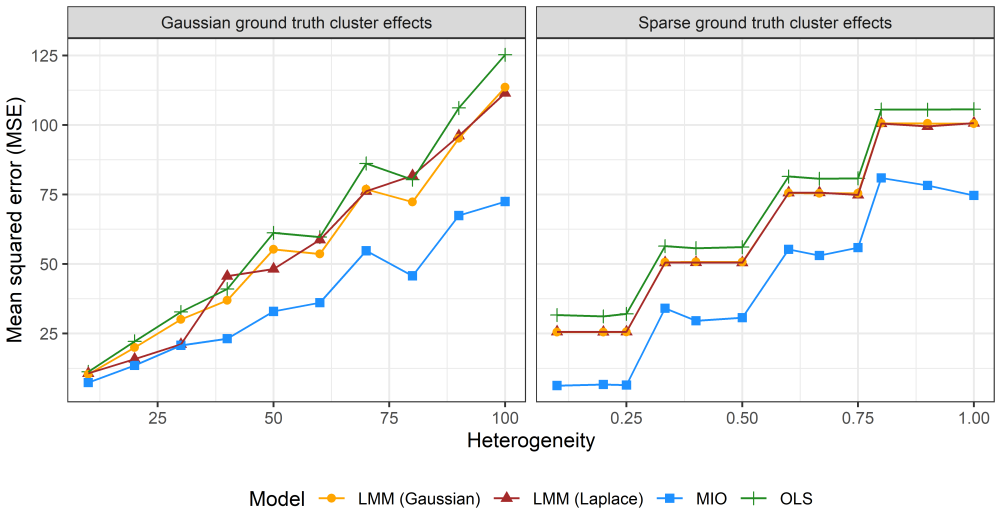

Fig B3: Predictive MSE (on the test set) of the four algorithms in a low dimensional setup (4 clusters, 10 covariates, 50 observations per cluster), where the cluster-effects are truly Gaussian (left) and truly sparse (right)

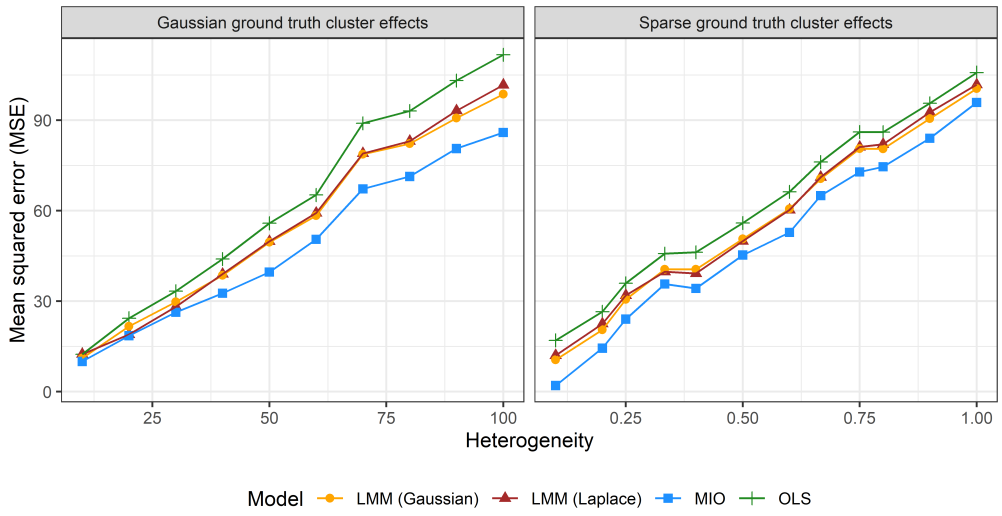

Fig B4: Predictive MSE (on the test set) of the four algorithms in a medium dimensional setup (10 clusters, 25 covariates, 50 observations per cluster), where the cluster-effects are truly Gaussian (left) and truly sparse (right)

C. Data example.

C.1. *Protein expression.* We utilize the Mouse Protein Expression Data from the UCI Machine Learning Repository. The dataset comprises 77 protein measurements obtained from the brains of 72 mice (Dua and Graff, 2017; Ahmed et al., 2015). Among these mice, 38 are controls and 34 are trisomic, afflicted with Down Syndrome. For each mouse, each protein is measured 15 times. It would be reasonable to postulate a form of cluster structure within the repeated observations for each mouse, irrespective of the different classes delineated in the dataset.

Higuera et al. (2015) probed the proteins that significantly differentiate between classes. Given that the class dependence remains consistent across observations of the same mouse, we can presume a level of cluster effect influencing the expression of these pivotal proteins. Consequently, to assess the efficacy of our algorithm against existing methods, we construct a linear regression problem using these significant proteins to discern if accounting for clustering enhances our predictive capacity.

Higuera et al. (2015) identifies 11 proteins as significant in differentiating between at least two classes of mice, inclusive of trisomic and control mice. These 11 proteins, representing a diverse array of biological pathways as per Table 3 of Higuera et al. (2015), will be the focus of our analysis. We aim to examine the effect of other proteins from various biological pathways on these significant proteins.

Employing complete cases ($n = 552$), we train a regression model on each of the aforementioned proteins, using the remaining 76 proteins as covariates, and evaluate predictive performance on a reserved test dataset. Figure C1 displays the predictive performances of the various methodologies.

It is noteworthy that both Gaussian and Laplace LMM approaches fail to converge in this setup, owing to complications in resolving singularities in data that emerge when making distributional assumptions. Despite this, we observe an improvement in predictive performance between Ordinary Least Squares (OLS) regression (no clusters assumed) and our cluster-informed MIO approach for 7 out of the 11 proteins. This suggests the presence of inherent cluster structure in the data that can be exploited for enhanced regression analysis. In this context, LMM is incapable of modeling this clustered nature, illustrating a limitation of distribution-based models; well-behaved data is a prerequisite for models like LMM to function effectively.

C.2. *Student data.* The dataset was obtained from UCI Machine Learning Repository . It features information about the academic performance of 382 Portuguese secondary school students, as explored in a study by Cortez and Silva (2008). The dataset encompasses 30 predictors, comprising demographic, social, and school-related factors such as parental education and study time. Continuous predictors were standardized, and dummy variables were introduced for categorical predictors, resulting in a total of 20 predictors. The outcome variables are the students' test scores in Mathematics and Portuguese.

In the main discussion of our paper, we contrasted the cluster-informed rankings yielded by the predictive methods with the rankings derived from raw scores. We observed that the MIO method identifies a distinct set of "exceptional" students compared to the raw score method. To validate that the algorithms exhibit reasonable cluster effects, we examine how well they recover the $\boldsymbol{\beta}$ vector. We adopt two strategies for this purpose. The first is to evaluate the population predictive Mean Squared Error (MSE) using the recovered $\boldsymbol{\beta}$ vector, and the second is to assess the pairwise concordance of each vector. The benchmark for comparison will be the vector resulting from Ordinary Least Squares (OLS) regression. Table C1 presents the predictive MSE for each method, which appear to be on a comparable scale. Figure C2

illustrates the pairwise correlations of the vectors. The evidence suggests that the examination of cluster effects is well warranted as all the algorithms are comparably efficacious in recovering the fixed effects.

TABLE C1

The predictive MSE of each regression method. Note that this is a population level prediction, and only leverages $\boldsymbol{\beta}$ values

	OLS	LME (Gaussian)	LME (Laplace)	MIO
MSE	10.45442	10.65313	10.81555	10.79221

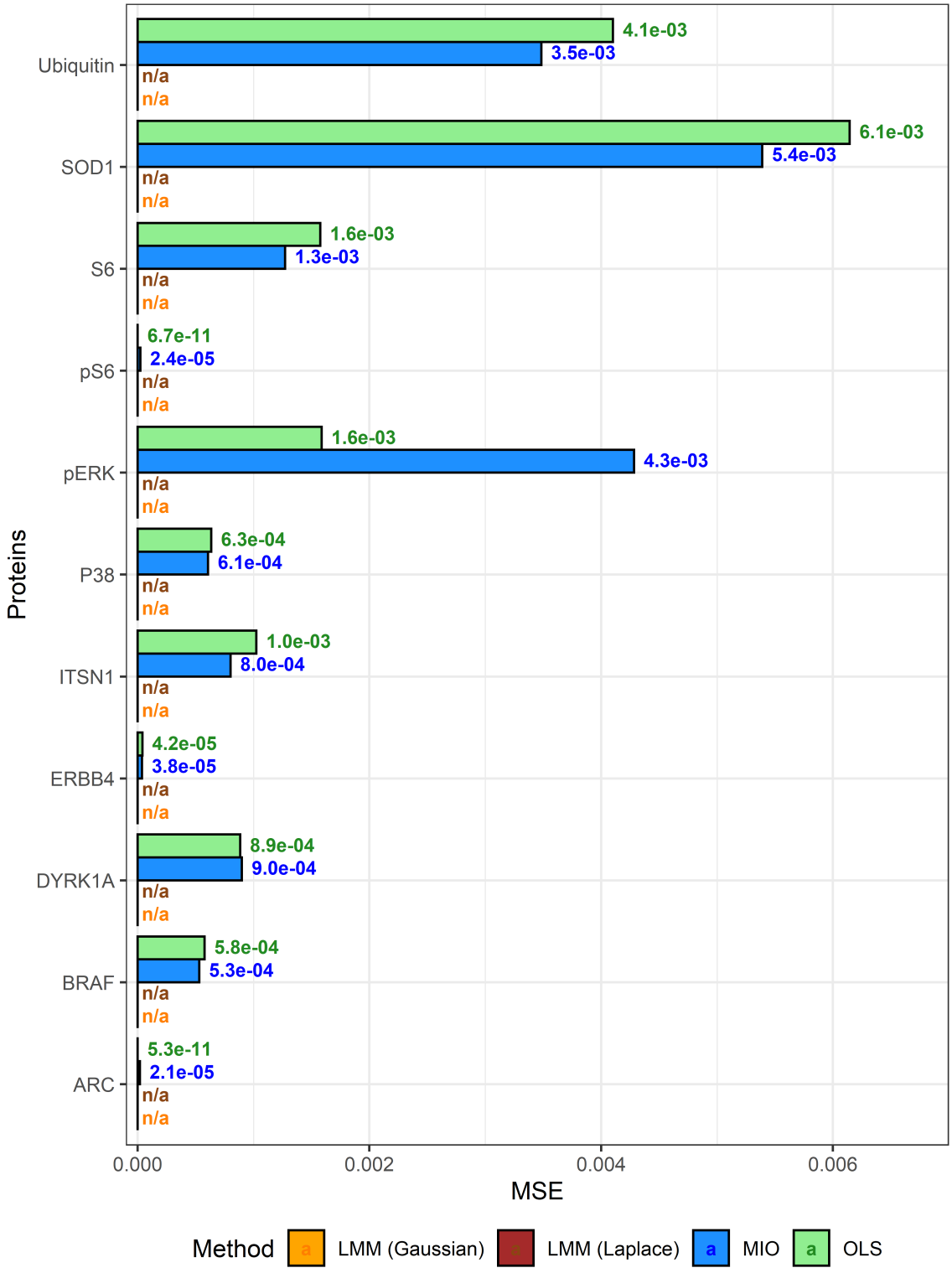

Fig C1: Predictive performance of MIO and OLS on significant proteins from Higuera et al. (2015). **LMM (Gaussian) and LMM (Laplace) do not produce any results due to convergence issues.**

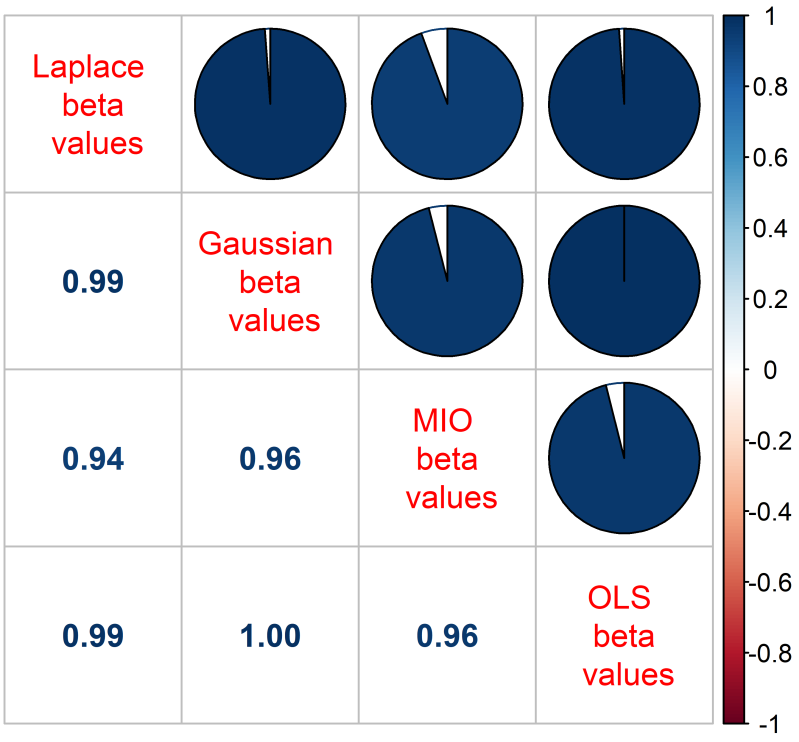

Fig C2: The pie charts on the upper-diagonal portion and values on the lower-diagonal portion represent the correlation of the β vectors between each pair of methods

REFERENCES

Ahmed MM, Dhanasekaran AR, Block A, Tong S, Costa ACS, Stasko M, Gardiner KJ (2015) Protein dynamics associated with failed and rescued learning in the ts65dn mouse model of down syndrome. *PLOS ONE* 10:e0119491.

Bertsimas D, King A, Mazumder R (2015) Best subset selection via a modern optimization lens arXiv:1507.03133 [math, stat].

Cortez P, Silva AMG (2008) Using data mining to predict secondary school student performance .

Dua D, Graff C (2017) UCI machine learning repository.

Higuera C, Gardiner KJ, Cios KJ (2015) Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PLOS ONE* 10:e0129126.