Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mixed-effects extreme gradient boosting for high-dimensional longitudinal data with continuous outcomes

Oyebayo Ridwan Olaniran

ridwan.olaniran@kcl.ac.uk

King's College London

Saheed Kunle Ajibade

Kwara State University

Samuel Adewale Aderoju

Kwara State University

Abdulhameed Bolakale

University of Ilorin

Nada MohammedSaeed Alharbi

Taibah University

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11734147/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Mixed-effects extreme gradient boosting for high-dimensional longitudinal data with continuous outcomes

Oyebayo Ridwan Olaniran[1,2,3*], Saheed Kunle Ajibade[2], Samuel Adewale Aderoju[2], Abdulhameed Bolakale[3], Nada MohammedSaeed Alharbi[4]

[1*]Department of Biostatistics and Health Informatics, King's College London, London, United Kingdom.
[2]Department of Statistics and Mathematical Sciences, Faculty of Pure and Applied Sciences, Kwara State University, PMB 1530, Malete, Kwara State, Nigeria..
[3]Department of Statistics, Faculty of Physical Sciences, University of Ilorin, PMB 1515, Ilorin, Kwara State, Nigeria..
[4]Department of Mathematics, Taibah University, Faculty of Science, Al-Madinah,42353, Al-Munawara, Saudi Arabia.

*Corresponding author(s). E-mail(s): ridwan.olaniran@kcl.ac.uk;

Abstract

Tree boosting accommodates high-dimensional predictors but typically treats repeated measurements as independent, overlooking within-subject dependence. Existing mixed-effects tree methods alternate between fitting a tree learner and predicting random effects. With a sufficiently flexible learner, this procedure can absorb subject-specific variation into the fixed-effect component and shrink the estimated random-effects covariance. We propose mixed-effects extreme gradient boosting (MEXGB) for continuous longitudinal outcomes. The method combines a relaxed lasso with second-order boosted trees, uses cross-fitted fixed-effect predictions to update the random effects and variance components, and selects model complexity using out-of-fold prediction error. We derive an oracle inequality for candidate selection, establish consistency of variance-component estimation under consistency of the fixed-effect estimator, and obtain prediction-error lower bounds for methods that ignore subject specific effects, misspecified linear mixed models, and tree-only fixed effects. MEXGB had the lowest mean

prediction error across 13 simulation scenarios with up to 2,000 predictors and ten competing methods. In the six main scenarios, it reduced mean squared error by 13–83% relative to mixed-effects gradient boosting and by 42–95% relative to extreme gradient boosting (XGBoost). In addition, MEXGB estimated the fixed effect more accurately than mixed-effects gradient boosting in every replicate. Across three applications, MEXGB was the most accurate method on one dataset and the most accurate mixed-effects method on a microbiome dataset with 1,879 predictors. The method is implemented in the R package `MEXGB`.

Keywords: Boosting, Cross-fitting, Linear mixed model, Longitudinal data, Random effects, Variable selection

1 Introduction

Longitudinal studies collect repeated measurements from the same individuals over time. Measurements from the same individual are typically correlated. Linear mixed-effects models (Laird and Ware 1982) account for this dependence through subject-specific random effects and predict additional measurements using best linear unbiased predictors (BLUPs) of those effects (Henderson 1975; Robinson 1991). Contemporary studies may also record hundreds or thousands of predictors at each visit, such as microbiome composition, gene-expression profiles, or wearable-sensor features. These predictors may affect the outcome through nonlinear relationships and interactions. A linear mean structure can therefore be restrictive, and an unpenalised model may not be identifiable when the number of predictors exceeds the number of measurements.

Tree ensembles, including random forests (Breiman 2001) and gradient boosting (Friedman 2001; Chen and Guestrin 2016), capture nonlinear relationships and interactions, perform implicit variable selection, and accommodate large numbers of predictors. However, standard applications to longitudinal data treat repeated measurements as independent. This limits prediction for individuals already observed during training, a task relevant to monitoring and personalised prediction. A predictor that does not use subject identity cannot explicitly incorporate the subject's estimated random effects, leaving subject-specific variation unaccounted for.

Mixed-effects tree methods combine these approaches by alternating between two steps: fitting a tree learner to outcomes after subtracting the current random-effect contribution, and updating the random effects through their BLUPs while re-estimating the variance components. This procedure underlies mixed-effects regression trees and random-effects expectation–maximisation (REEM) trees (Hajjem et al. 2011; Sela and Simonoff 2012), mixed-effects random forests (Hajjem et al. 2014; Capitaine et al. 2021), and mixed-effects gradient boosting (MEGB; Olaniran et al. 2025). GPBoost instead combines tree boosting with grouped random effects estimated jointly through marginal likelihood (Sigrist 2022, 2023). Component-wise boosting for mixed models has also been reformulated to reduce competition between fixed and random effects during variable selection (Griesbach et al. 2021; Knieper et al. 2025).

The alternating estimation scheme is susceptible to overfitting when the fixed-effects learner exhibits high model flexibility. For instance, a large tree ensemble with shallow terminal nodes can approximate the training responses to such an extent that subject-specific random-effect variation is inadvertently absorbed into the fixed-effects fitted values. Consequently, the residuals subsequently used to estimate the random effects contain limited random-effect signal, and the corresponding covariance estimates may progressively collapse toward zero over successive iterations. In this regime, the resulting method may effectively reduce to the underlying base learner, providing little to no contribution from random effects. Although constraints on ensemble complexity, bagging strategies, and joint estimation approaches can attenuate this behavior, they do not directly resolve the fundamental reliance on in-sample fixed-effects predictions within the alternating updates.

Motivated by these limitations, we propose mixed-effects extreme gradient boosting (MEXGB) for continuous longitudinal outcomes. The proposed method comprises four components.

1. A partially linear fixed-effects specification of the form $f(\mathbf{x}) = a + \mathbf{x}^\top \boldsymbol{\beta} + T(\mathbf{x})$. The sparse linear component $\mathbf{x}^\top \boldsymbol{\beta}$ is estimated using the relaxed lasso (Meinshausen 2007), whereas the residual nonlinear structure is captured by second-order (Newton) boosted decision trees $T(\mathbf{x})$ (Chen and Guestrin 2016), with $\mathbf{x}$ denoting the predictor vector. Incorporating an explicit linear term mitigates the need for tree-based learners to approximate linear associations via piecewise-constant step functions.
2. A cross-fitted expectation step. Random effects and variance components are updated using out-of-fold fixed-effect predictions from V fold models. Each model excludes the measurements it predicts while retaining other measurements from the same subjects. This approach is intended to reduce the absorption of random-effect variation into the fixed-effect fit.
3. Data-driven selection of model complexity. The number of trees, the screened predictor set, and the inclusion of the linear component are selected using out-of-fold loss. For each held-out measurement, the random effect is predicted using the subject's measurements in the other folds. Under some regularity conditions, candidate selection satisfies an oracle inequality with a remainder of order $\{\log(\text{number of candidates})/n\}^{1/2}$, where n is the number of subjects.
4. A cross-validated covariance check. The EM estimate of the random-effects covariance is retained unless a shrunken estimate improves held-out prediction according to a prespecified decision rule.

We also derive lower bounds for predictors that ignore subject identity, misspecified linear mixed models, and piecewise-constant approximations to linear signals. Under the stated regularity conditions, variance-component estimation is consistent when the fixed-effect estimator is consistent in empirical L_2.

The empirical evaluation extends the designs used in the published MEGB and longitudinal random forest studies (Capitaine et al. 2021; Olaniran et al. 2025). It uses a larger sample, linear and nonlinear fixed effects, and up to 2000 predictors. A sensitivity analysis varies predictor correlation, residual serial correlation, and noise

level. We compare MEXGB with ten methods across varying scenarios. A separate estimation study examines recovery of the variance components, random effects, and fixed effect, and includes two reduced versions of MEXGB to investigate the roles of the linear component and the cross-fitted expectation step. Three applications illustrate the method, including a gut-microbiome dataset with 1879 predictors. MEXGB is implemented in the R package `MEXGB`.

Section 2 reviews related work. Section 3 presents the model and estimation algorithm, and Section 4 develops the theoretical results. Sections 5 and 6 report the simulations and applications, respectively. Section 7 discusses the findings, limitations, and extensions, and Section 8 concludes. Appendix A contains the proofs, Appendix B provides a package tutorial, and Appendix C presents additional results.

2 Related work

Linear and penalised mixed models.

Comprehensive treatments of linear mixed models, including maximum-likelihood and restricted maximum-likelihood estimation, are provided by Verbeke and Molenberghs (2000), Diggle et al. (2002), and Fitzmaurice et al. (2011). For repeated-measures settings, Laird et al. (1987) introduced an EM algorithm; its typically linear, and in some instances slow, rate of convergence is discussed by Wu (1983) and McLachlan and Krishnan (2008). Methodological developments for high-dimensional regimes commonly incorporate penalisation, including the ℓ_1-penalised linear mixed model of Schelldorfer et al. (2011), the generalised mixed-model extension implemented in `glmmLasso` (Groll and Tutz 2014), and the joint penalisation of fixed and random effects proposed by Yang et al. (2024). Notably, these contributions maintain a linear predictor with respect to the fixed-effects component.

Mixed-effects trees and forests.

Hajjem et al. (2011) and Sela and Simonoff (2012) propose iterative procedures that alternate between fitting a regression tree to the pseudo-response $\mathbf{y}_i - \mathbf{Z}_i\hat{\mathbf{b}}_i$, where $\mathbf{y}_i$ denotes the response vector and $\mathbf{Z}_i\hat{\mathbf{b}}_i$ is the random-effects contribution obtained using the BLUP of $\mathbf{b}_i$; Fu and Simonoff (2015) implement an analogous scheme using conditional inference trees. Mixed-effects random forests (MERF; Hajjem et al. 2014) generalise this approach by replacing the single tree with a random forest, and Capitaine et al. (2021) extend MERF to high-dimensional longitudinal settings with an optional stochastic-process component (`LongituRF`, REEMforest). Extensions to non-Gaussian outcomes have been developed by Hajjem et al. (2017), Speiser et al. (2019), and Pellagatti et al. (2021), while Ngufor et al. (2019) combine multiple learners with random effects. Alternative formulations include the model-based trees of Fokkema et al. (2018), in which terminal nodes are parameterised by generalised linear mixed models, and the Bayesian additive regression tree approach of Tan et al. (2018), which incorporates a random intercept. For broader context, Hu and Szymczak (2023) provide a review of random-forest methods for longitudinal data, and Fu et al. (2026) propose forests whose nodes summarise sparse and irregular trajectories.

Boosting with random effects.

Likelihood-based boosting approaches for (additive) mixed models were introduced by Tutz and Groll (2010) and Groll and Tutz (2012), while Griesbach et al. (2021) proposed gradient boosting for linear mixed models incorporating a correction for cluster-constant covariates. Building on these contributions, Knieper et al. (2025) extend the framework to additive mixed models in `mermboost`, where the random effects and their covariance structure are treated as nuisance parameters and updated at each iteration. This strategy mitigates competition between fixed and random effects during component-wise model selection. GPBoost (Sigrist 2022, 2023) applies boosting to tree-based fixed-effect components and estimates grouped random effects or Gaussian processes via marginal likelihood. MEGB (Olaniran et al. 2025) embeds first-order gradient boosting with a fixed number of trees within the EM/BLUP iteration scheme of MERF, and demonstrated improved predictive accuracy over MERF and REEMforest in the high-dimensional scenarios considered. In related work, Messer and Schmid (2024) coupled XGBoost with a linear mixed model through EM-based alternation for small area estimation; Cho and Mueller (2026) integrated XGBoost with crossed random intercepts and introduced group-aware cross-validation and permutation-based importance measures; and Toivonen et al. (2026) reported that boosted mixed models yielded more accurate predictions of forest age than linear mixed models. Finally, GBMixed (Prevett et al. 2025) jointly boosts the conditional mean and variance components of a Gaussian mixed model using gradients of the marginal log-likelihood.

Contribution of this work.

MEXGB is situated within the EM/BLUP class of methods introduced by Hajjem et al. (2014), Capitaine et al. (2021), and Olaniran et al. (2025), while differing from these approaches in three key respects. First, the fixed-effect component is specified as partially linear, rather than being represented exclusively by tree-based learners. Second, in the expectation step, MEXGB relies on cross-fitted predictions to mitigate the influence of in-sample overfitting on the estimation of the random effects. Third, model complexity is tuned using an out-of-fold criterion tailored to the prediction of additional measurements for subjects that are observed during training. Concretely, measurements are partitioned within each subject, and for each held-out measurement the corresponding random effect is predicted using the remaining measurements from that subject. This cross-fitting strategy instantiates the general principle of decoupling flexible nuisance estimation from the estimation of target parameters (Chernozhukov et al. 2018); in the present setting, the targets are the fixed effects and the random effects, along with their covariance structure. In contrast to GPBoost, MEXGB does not estimate random effects jointly with the tree ensemble via marginal-likelihood optimization; instead, it performs Gaussian working-model updates and determines the number of trees by minimizing out-of-fold prediction error.

3 Methods

3.1 Model specification

Consider a sample of n independent subjects indexed by $i \in \{1, \dots, n\}$. Subject i provides n_i repeated measurements indexed by $j \in \{1, \dots, n_i\}$, with total sample size $N = \sum_{i=1}^{n} n_i$. For each observation (i, j), let y_{ij} denote a continuous response, $\mathbf{x}_{ij} \in \mathbb{R}^p$ a vector of p predictors, and $\mathbf{z}_{ij} \in \mathbb{R}^q$ a q-dimensional random-effects design vector. We allow p to be substantially larger than N, whereas q is assumed to be small. In the applications considered here, $\mathbf{z}_{ij} = (1, t_{ij})^\top$ encodes a random intercept and a random slope with respect to measurement time t_{ij}. For subject i, we collect the responses and design matrices as $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^\top$, $\mathbf{X}_i = (\mathbf{x}_{i1}^\top; \dots; \mathbf{x}_{in_i}^\top) \in \mathbb{R}^{n_i \times p}$, and $\mathbf{Z}_i = (\mathbf{z}_{i1}^\top; \dots; \mathbf{z}_{in_i}^\top) \in \mathbb{R}^{n_i \times q}$.

The MEXGB model is specified as

$$y_{ij} = f(\mathbf{x}_{ij}) + \mathbf{z}_{ij}^\top \boldsymbol{\beta}_z + \mathbf{z}_{ij}^\top \mathbf{b}_i + \varepsilon_{ij}, \qquad \mathbf{b}_i \overset{\text{iid}}{\sim} \mathcal{N}_q(\mathbf{0}, \mathbf{D}), \qquad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \tag{1}$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is the fixed-effect regression function, $\boldsymbol{\beta}_z \in \mathbb{R}^q$ is the population-level coefficient vector associated with $\mathbf{z}_{ij}$, $\mathbf{b}_i \in \mathbb{R}^q$ is a subject-specific random-effect vector with covariance matrix $\mathbf{D} \in \mathbb{R}^{q \times q}$, and $\varepsilon_{ij} \in \mathbb{R}$ is an idiosyncratic error with variance $\sigma^2 > 0$; $\mathcal{N}_q(\mathbf{0}, \mathbf{D})$ denotes the q-variate Gaussian distribution with mean $\mathbf{0}$ and covariance $\mathbf{D}$, and $\mathcal{N}(0, \sigma^2)$ denotes the univariate Gaussian with mean 0 and variance σ^2. The random effects $\mathbf{b}_i$ are assumed independent of the residual errors ε_{ij}.

The fixed-effect component f is assumed to have a partially linear decomposition,

$$f(\mathbf{x}) = a + \mathbf{x}^\top \boldsymbol{\beta} + T(\mathbf{x}), \qquad T(\mathbf{x}) = \sum_{m=1}^{M} \nu\, t_m(\mathbf{x}), \qquad t_m(\mathbf{x}) = \sum_{\ell=1}^{L_m} w_{m\ell} \mathbb{1}\{\mathbf{x} \in R_{m\ell}\}. \tag{2}$$

Here, $a \in \mathbb{R}$ is an intercept; $\boldsymbol{\beta} \in \mathbb{R}^p$ is a sparse vector of linear coefficients; and $T(\mathbf{x})$ is an additive tree-based component comprising $M \in \mathbb{N}$ regression trees. For tree $m \in \{1, \dots, M\}$, $t_m(\cdot)$ is a piecewise-constant function defined by a partition of the predictor space into L_m terminal regions (leaves) $\{R_{m\ell}\}_{\ell=1}^{L_m}$ (with tree depth at most d); $w_{m\ell} \in \mathbb{R}$ is the leaf weight for region $R_{m\ell}$; $\mathbb{1}\{\cdot\}$ is the indicator function; and $\nu \in (0, 1]$ is the learning-rate (shrinkage) parameter.

Special cases are obtained by omitting either component of f: setting $\boldsymbol{\beta} = \mathbf{0}$ yields a purely tree-based fixed effect as in MERF and MEGB, whereas setting $M = 0$ reduces the specification to a penalised linear mixed model.

Integrating out $\mathbf{b}_i$ gives $\mathbf{y}_i \sim \mathcal{N}_{n_i}(\mathbf{f}_i + \mathbf{Z}_i \boldsymbol{\beta}_z, \mathbf{V}_i)$ with $\mathbf{f}_i = (f(\mathbf{x}_{i1}), \dots, f(\mathbf{x}_{in_i}))^\top$ and $\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^\top + \sigma^2 \mathbf{I}_{n_i}$, and the log-likelihood is

$$\ell(f, \boldsymbol{\beta}_z, \mathbf{D}, \sigma^2) = -\frac{1}{2} \sum_{i=1}^{n} \Big\{ \log |\mathbf{V}_i| + (\mathbf{y}_i - \mathbf{f}_i - \mathbf{Z}_i \boldsymbol{\beta}_z)^\top \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{f}_i - \mathbf{Z}_i \boldsymbol{\beta}_z) \Big\} + \text{const.} \tag{3}$$

Assumption 1 (i) $\mathbb{E}\mathbf{b}_i = \mathbf{0}$, and $\mathbf{b}_i$ is independent of $\{\mathbf{x}_{ij}, \mathbf{z}_{ij}\}_j$ and of the errors. (ii) $\mathbf{Z}_i$ has full column rank for a positive fraction of the subjects, and the columns of $\mathbf{z}$ other than the constant are not functions of $\mathbf{x}$.

Under Assumption 1(i), f represents the population-level relationship between the predictors and the outcome, with the random effects independent of the predictors. When time is also a predictor, only the sum $f(\mathbf{x})+\mathbf{z}^\top\boldsymbol{\beta}_z$ is identified, which is sufficient for prediction. The parameters $(\boldsymbol{\beta}_z, \{\mathbf{b}_i\})$ and $(\boldsymbol{\beta}_z+\mathbf{m}, \{\mathbf{b}_i-\mathbf{m}\})$ give the same linear predictor for every $\mathbf{m} \in \mathbb{R}^q$; the constraint $\mathbb{E}\mathbf{b}_i = \mathbf{0}$ fixes $\mathbf{m}$. Without $\boldsymbol{\beta}_z$, a population time trend that f does not capture could instead enter the mean of the predicted random effects and inflate the estimate of $\mathbf{D}$.

3.2 The MEXGB algorithm

Rather than maximising (3) directly over an unrestricted function f, MEXGB uses a scheme based on expectation–conditional maximisation (Meng and Rubin 1993) in which the conditional maximisation over f is replaced by a boosting fit. Algorithm 1 summarises the estimation procedure; Section 3.3 explains each step.

3.3 Details of the algorithm

Initialisation.

The random effects are initialised at zero, so the first fixed-effect fit uses the observed responses without a random-effect adjustment. If the predictors explain part of the differences between subjects, the fixed effect can capture this variation before the random effects are updated. The folds are formed within subjects: the measurements of each subject are spread as evenly as possible over the V folds, so every subject is represented in every training complement $\mathcal{F}_v^c$.

Fixed-effect update: linear component.

Conditional on the offsets, the linear component is obtained by solving the lasso optimisation problem (Tibshirani 1996)

$$(\hat{a}_\lambda, \hat{\boldsymbol{\beta}}_\lambda) = \arg\min_{a,\boldsymbol{\beta}} \Big\{ \frac{1}{2N} \sum_{i,j} (y_{ij} - o_{ij} - a - \mathbf{x}_{ij}^\top \boldsymbol{\beta})^2 + \lambda_{\mathrm{L}} \|\boldsymbol{\beta}\|_1 \Big\}, \tag{4}$$

over a grid (path) of candidate values of λ_{L}, using coordinate descent (Friedman et al. 2010). For $\gamma_{\mathrm{R}} \in \{0, \frac{1}{2}, 1\}$, we form a relaxed estimator as the convex combination $\gamma_{\mathrm{R}}(\hat{a}_\lambda, \hat{\boldsymbol{\beta}}_\lambda) + (1-\gamma_{\mathrm{R}})(\tilde{a}_\lambda, \tilde{\boldsymbol{\beta}}_\lambda)$, where $(\tilde{a}_\lambda, \tilde{\boldsymbol{\beta}}_\lambda)$ denotes the least-squares refit restricted to the active set of $\hat{\boldsymbol{\beta}}_\lambda$ (Meinshausen 2007; Hastie et al. 2020). The hyperparameter pair $(\lambda_{\mathrm{L}}, \gamma_{\mathrm{R}})$ is selected by minimising the out-of-fold loss in (10). The least-squares refit is intended to mitigate shrinkage induced by the ℓ_1 penalty. To reduce extrapolation at atypical predictor configurations, predictions from the linear component are truncated to the range of fitted values observed on the training data. This truncation cannot increase the prediction error whenever the target value lies within that range.

Algorithm 1 MEXGB estimation procedure

Require: data $\{(\mathbf{x}_{ij}, \mathbf{z}_{ij}, y_{ij})\}$; learning rate ν, depth d, minimum leaf size, leaf penalty λ, maximum number of trees $M_{\max}$; number of folds V; maximum inner iterations L_{in} and tolerance ϵ_{in}; maximum outer iterations $T_{\max}$, tolerance δ and patience.

$\hat{\mathbf{b}}_i \leftarrow \mathbf{0}$, $\boldsymbol{\beta}_z \leftarrow \mathbf{0}$, $\widehat{\mathbf{D}} \leftarrow \mathbf{I}_q$, $\hat{\sigma}^2 \leftarrow 1$; split the measurements of every subject into V folds $\mathcal{F}_1, \ldots, \mathcal{F}_V$.

for $t = 1, \ldots, T_{\max}$ **do**

a) Offsets. Training offsets $o_{ij} = \mathbf{z}_{ij}^\top(\boldsymbol{\beta}_z + \hat{\mathbf{b}}_i)$. Held-out offsets $\tilde{o}_{ij} = 0$ if $t = 1$, otherwise $\tilde{o}_{ij} = \mathbf{z}_{ij}^\top\{\boldsymbol{\beta}_z + \hat{\mathbf{b}}_i^{(-v)}\}$ for $(i,j) \in \mathcal{F}_v$, by (11).

b) Fixed effect. Fit a relaxed lasso to $y - o$ using $\mathbf{x}$, on the full data and on each training complement $\mathcal{F}_v^c$; choose its penalty by the out-of-fold loss (10).

Grow Newton trees on the residuals of the linear component (fold models in chunks); choose the number of trees $\hat{M}$ by (10).

If $t \leq 2$: screen the predictors and decide on the linear component (Section 3.3).

Full-data fit $\hat{f}^{\text{full}}$ with $\lceil \hat{M}V/(V-1) \rceil$ trees; cross-fitted fit $\hat{f}^{\text{cf}}(\mathbf{x}_{ij}) = \hat{f}^{(-v)}(\mathbf{x}_{ij})$, $(i,j) \in \mathcal{F}_v$.

c) Random effects. Repeat up to L_{in} times, until the relative change of $(\widehat{\mathbf{D}}, \hat{\sigma}^2)$ is below ϵ_{in}: BLUPs (7) of $\mathbf{y}_i - \hat{f}_i^{\text{cf}} - \mathbf{Z}_i\boldsymbol{\beta}_z$; transfer their mean to $\boldsymbol{\beta}_z$; update $(\widehat{\mathbf{D}}, \hat{\sigma}^2)$ by (8).

d) Covariance check. Compute the out-of-fold error $\mathcal{E}(c)$ of (13) for $c \in \{0, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\}$; if some $c < 1$ improves on $c = 1$ according to (12), set $\widehat{\mathbf{D}} \leftarrow c^\star\widehat{\mathbf{D}}$ and recompute the BLUPs.

e) Criterion. $\kappa^{(t)} = \mathcal{E}(c^\star)/N$; store the state if $\kappa^{(t)}$ is the smallest so far.

Stop if $|\kappa^{(t)} - \kappa^{(t-1)}| < \delta\kappa^{(t-1)}$ or κ has not improved for *patience* iterations.

end for

Ensure: Return the stored state: $\hat{f}^{\text{full}}$ (linear component, trees, selected predictors), $\boldsymbol{\beta}_z$, $\{\hat{\mathbf{b}}_i\}$, $\widehat{\mathbf{D}}$, $\hat{\sigma}^2$.

Fixed-effect update: Newton trees (XGBoost).

At boosting step m, we fit a regression tree to the remaining nonlinear signal in y_{ij} after accounting for the current offset o_{ij} and the estimated linear component $\hat{a} + \mathbf{x}_{ij}^\top\hat{\boldsymbol{\beta}}$. In XGBoost terminology, the quantity $o_{ij} + \hat{a} + \mathbf{x}_{ij}^\top\hat{\boldsymbol{\beta}}$ is the *base margin* (the current prediction before adding the new tree).

Suppose that we let the current (working) linear predictor be η_{ij}, and let $L(y, \eta)$ denote a differentiable pointwise loss. Define the first and second derivatives with respect to η at the current predictor, $g_{ij} = \partial L(y_{ij}, \eta_{ij})/\partial\eta$ and $h_{ij} = \partial^2 L(y_{ij}, \eta_{ij})/\partial\eta^2$. The m-th tree is a function $t(\cdot)$ that maps $\mathbf{x}$ to a leaf weight: if $q(\mathbf{x}) \in \{1, \ldots, L_t\}$ denotes the leaf index assigned by the tree structure (with L_t leaves), then $t(\mathbf{x}) = w_{q(\mathbf{x})}$ for leaf weights $\{w_\ell\}_{\ell=1}^{L_t}$. The tree and its leaf weights are chosen to minimise the

second-order approximation of Chen and Guestrin (2016),

$$\sum_{i,j}\Big\{g_{ij}t(\mathbf{x}_{ij}) + \tfrac{1}{2}h_{ij}t(\mathbf{x}_{ij})^2\Big\} + \gamma L_t + \frac{\lambda}{2}\sum_{\ell=1}^{L_t} w_\ell^2, \tag{5}$$

where $\gamma \geq 0$ penalises the number of leaves (tree complexity) and $\lambda \geq 0$ is the ℓ_2 penalty on leaf weights. For a fixed tree structure, the optimal weight in leaf ℓ is $w_\ell^\star = -G_\ell/(H_\ell + \lambda)$, where $G_\ell = \sum_{(i,j):\, q(\mathbf{x}_{ij})=\ell} g_{ij}$ and $H_\ell = \sum_{(i,j):\, q(\mathbf{x}_{ij})=\ell} h_{ij}$. For a candidate split that partitions a parent leaf into left and right children (with aggregated derivatives $(G_\mathrm{L}, H_\mathrm{L})$ and $(G_\mathrm{R}, H_\mathrm{R})$), the reduction in the objective is

$$\text{Gain} = \frac{1}{2}\Big\{\frac{G_\mathrm{L}^2}{H_\mathrm{L}+\lambda} + \frac{G_\mathrm{R}^2}{H_\mathrm{R}+\lambda} - \frac{(G_\mathrm{L}+G_\mathrm{R})^2}{H_\mathrm{L}+H_\mathrm{R}+\lambda}\Big\} - \gamma. \tag{6}$$

For the squared-error loss $L(y,\eta) = \frac{1}{2}(y-\eta)^2$, we have $g_{ij} = \eta_{ij} - y_{ij} = -(y_{ij} - \eta_{ij})$ (minus the current residual) and $h_{ij} = 1$. In this case, $w_\ell^\star$ equals the mean residual in leaf ℓ multiplied by $|\ell|/(|\ell| + \lambda)$, where $|\ell|$ is the number of observations assigned to that leaf; thus λ shrinks updates more strongly in small leaves, stabilising estimation even when a leaf contains only a few observations (e.g., two measurements). The total number of trees can then be increased to offset the smaller per-tree updates.

Random-effects update.

For a fixed estimate of the mean, define the residual vector $\mathbf{r}_i = \mathbf{y}_i - \hat{f}_i^{\mathrm{cf}} - \mathbf{Z}_i\boldsymbol{\beta}_z$. The Gaussian working-model expectation step uses

$$\hat{\mathbf{b}}_i = \widehat{\mathbf{D}}\mathbf{Z}_i^\top \mathbf{V}_i^{-1}\mathbf{r}_i, \qquad \mathbf{C}_i = \widehat{\mathbf{D}} - \widehat{\mathbf{D}}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}\mathbf{Z}_i\widehat{\mathbf{D}} = \big(\widehat{\mathbf{D}}^{-1} + \hat{\sigma}^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\big)^{-1} \tag{7}$$

as the conditional mean and covariance of $\mathbf{b}_i$ given $\mathbf{y}_i$. After each update, the mean $\bar{\mathbf{b}}$ of the $\hat{\mathbf{b}}_i$ is transferred to the population coefficients, $\boldsymbol{\beta}_z \leftarrow \boldsymbol{\beta}_z + \bar{\mathbf{b}}$ and $\hat{\mathbf{b}}_i \leftarrow \hat{\mathbf{b}}_i - \bar{\mathbf{b}}$, which leaves every linear predictor unchanged. The variance components are updated by the EM formulas

$$\widehat{\mathbf{D}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^\top + \mathbf{C}_i), \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n}\Big\{\|\mathbf{r}_i - \mathbf{Z}_i\hat{\mathbf{b}}_i\|^2 + \mathrm{tr}(\mathbf{Z}_i\mathbf{C}_i\mathbf{Z}_i^\top)\Big\}. \tag{8}$$

These three updates are repeated up to L_{in} times, with the fixed-effect predictions held at $\hat{f}^{\mathrm{cf}}$. For a fixed offset the iteration is the EM algorithm of the linear mixed model, which increases the marginal likelihood at every step. Convergence is linear and can be slow when the subject means have low reliability; repeating these updates for each boosting fit adds relatively little computational cost compared with the boosting fit itself. Each pass over the subjects costs $O(\sum_i n_i q^2 + nq^3)$.

Cross-fitted expectation step.

The expectation step uses the cross-fitted prediction instead of the full-data fit $\hat{f}^{\text{full}}$:

$$\hat{f}^{\text{cf}}(\mathbf{x}_{ij}) = \hat{f}^{(-v)}(\mathbf{x}_{ij}), \qquad (i,j) \in \mathcal{F}_v, \tag{9}$$

Here, $\hat{f}^{(-v)}$ is fitted on $\mathcal{F}_v^c$ using the selected linear-component specification, number of trees, and predictor set. The full-data fit is used only for prediction. Section 4.1 explains the motivation: a sufficiently flexible in-sample fit can depend heavily on y_{ij} itself, removing random-effect variation from the residuals and causing $\widehat{\mathbf{D}}$ to shrink across iterations. Under the assumptions of the cross-fitted smoother analysis, excluding y_{ij} retains the random-effect contribution in the residual up to a fraction of order n_i/N. Because every subject appears in every fold model, $\hat{f}^{\text{cf}}(\mathbf{x}_{ij})$ is an out-of-sample prediction of f at a measurement of a subject seen in training, which matches the intended prediction task.

Out-of-fold choices.

The fold models are also used to choose the complexity of the fixed effect. For a candidate fixed-effect specification $\mathfrak{f}$ (defined by a linear-component penalty, number of trees, or predictor set), with fold fits $\hat{\mathfrak{f}}^{(-v)}$, the out-of-fold loss is

$$\widehat{\mathcal{L}}^{\text{oof}}(\mathfrak{f}) = \sum_{v=1}^{V} \sum_{(i,j)\in\mathcal{F}_v} \left\{ y_{ij} - \tilde{o}_{ij} - \hat{\mathfrak{f}}^{(-v)}(\mathbf{x}_{ij}) \right\}^2. \tag{10}$$

The offset $\tilde{o}_{ij}$ of a held-out measurement must exclude y_{ij}; otherwise, the loss is biased in favour of the random effects. From the second outer iteration onwards, it uses the random-effect contribution predicted from the subject's measurements outside $\mathcal{F}_v$,

$$\hat{\mathbf{b}}_i^{(-v)}(\mathbf{D}) = \mathbf{D}\mathbf{Z}_{i,-v}^{\top}\left(\mathbf{Z}_{i,-v}\mathbf{D}\mathbf{Z}_{i,-v}^{\top} + \hat{\sigma}^2\mathbf{I}\right)^{-1}\left(\mathbf{y}_{i,-v} - \hat{f}_{i,-v}^{\text{cf}} - \mathbf{Z}_{i,-v}\boldsymbol{\beta}_z\right), \tag{11}$$

the predictor that is used for a new measurement of subject i given the subject's other measurements. At the first iteration the random effects are zero and the offsets are zero.

Number of trees. The V fold models are grown in chunks of geometrically increasing size and the out-of-fold loss is recorded every ten trees, including $M = 0$. Growth stops after 200 additional trees without an improvement in the current minimum, or at $M_{\max} = 1000$. The selected value $\hat{M}$ minimises the recorded loss. The full-data model, fitted on $V/(V-1)$ times more measurements, uses $\lceil \hat{M}V/(V-1) \rceil$ trees.

Screening. When p is large, splits on inactive predictors can increase variance. The largest spurious gain among r inactive predictors grows like $\sigma^2 \log r$. During each of the first two outer iterations, each fold model ranks predictors by total split gain and is refitted using its own top K predictors, for $K \in \mathcal{K}$, and the screened sets are compared with the set of all predictors by (10). When the ranking and offsets for fold v use only $\mathcal{F}_v^c$, the comparison is unbiased for the corresponding fold risks.

Linear component. If $p \leq \sqrt{N}$ the linear component is always included. Otherwise, models with and without the linear component are fitted during the first two outer iterations, and the specification with the smaller out-of-fold loss is retained.

Decision rule for departures from the default. The screened set and the shrinkage of $\mathbf{D}$ described next replace the default (all predictors; the EM estimate of $\mathbf{D}$) only if they improve the held-out loss according to the following rule. Let d_{ij} denote the difference between the held-out squared errors of the default and the alternative. The rule is

$$z = \frac{\bar{d}}{s_d/\sqrt{N}} > 2, \qquad \bar{d} = \frac{1}{N}\sum_{i,j} d_{ij}, \qquad s_d^2 = \frac{1}{N-1}\sum_{i,j}(d_{ij} - \bar{d})^2. \tag{12}$$

The paired statistic accounts for the similarity of predictions from competing candidates. However, $s_d/\sqrt{N}$ does not account for within-subject dependence and may underestimate the standard error of $\bar{d}$. The threshold should therefore be interpreted as a heuristic decision rule, rather than as a calibrated significance test.

Covariance check.

After each expectation step the EM estimate is compared with shrunken versions through the out-of-fold squared error

$$\mathcal{E}(c) = \sum_{v=1}^{V} \sum_{(i,j)\in\mathcal{F}_v} \left\{ y_{ij} - \hat{f}^{\mathrm{cf}}(\mathbf{x}_{ij}) - \mathbf{z}_{ij}^\top \boldsymbol{\beta}_z - \mathbf{z}_{ij}^\top \hat{\mathbf{b}}_i^{(-v)}(c\widehat{\mathbf{D}}) \right\}^2 \tag{13}$$

for $c \in \{0, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\}$. If the minimiser $c^\star < 1$ improves on $c = 1$ according to (12), $\widehat{\mathbf{D}}$ is replaced by $c^\star\widehat{\mathbf{D}}$. Under the model, prediction using the true covariance is optimal. The check therefore retains the EM estimate unless a smaller covariance improves held-out prediction according to the decision rule. Such an improvement may occur when predictors explain some between-subject variation that the current fixed-effect fit has not captured. Shrinking $\mathbf{D}$ allows more of that variation to enter the next fixed-effect update.

Stopping criteria.

Because the cross-fitted updates do not maximise a single common objective, the outer iterations need not increase the likelihood. MEXGB therefore monitors the out-of-fold mean squared error $\kappa^{(t)} = \mathcal{E}(c^\star)/N$, which estimates the error of predicting new measurements of the training subjects, and retains the state with the smallest recorded value. It stops when the relative change is below δ, after five iterations without improvement, or after $T_{\max}$ iterations. The marginal log-likelihood (3) at $\hat{f}^{\mathrm{cf}}$ is recorded as a diagnostic.

Prediction.

Algorithm 2 describes prediction. For a subject observed during training, prediction uses the estimated conditional mean given the subject's training measurements. For a

Algorithm 2 Prediction with a fitted MEXGB model

Require: fitted model; new measurement $(\mathbf{x}^*, \mathbf{z}^*)$ of subject s.

1: $\hat{f}^* \leftarrow \Pi\big(\hat{a} + \mathbf{x}^{*\top}\hat{\boldsymbol{\beta}}\big) + \sum_{m \le M_{\text{full}}} \nu t_m(\mathbf{x}^*)$, where Π projects onto the training range of the linear component.

2: **if** s was seen in training **then** $\hat{y}^* \leftarrow \hat{f}^* + \mathbf{z}^{*\top}(\boldsymbol{\beta}_z + \hat{\mathbf{b}}_s)$ **else** $\hat{y}^* \leftarrow \hat{f}^* + \mathbf{z}^{*\top}\boldsymbol{\beta}_z$.

Ensure: $\hat{y}^*$.

new subject, the random effect is set to its population mean of zero. When additional measurements become available for a previously observed subject, the random effect can be updated by (7) with the fixed effect and variance components held fixed, without refitting the trees.

Variable importance.

The importance of predictor k combines the two components of the fixed effect,

$$I_k = \frac{1}{N} \sum_{i,j} \big| \phi_k(\mathbf{x}_{ij}) + \hat{\beta}_k (x_{ijk} - \bar{x}_k) \big|, \tag{14}$$

normalised to sum to 100%, where ϕ_k is the SHAP contribution of predictor k to the tree component (Lundberg and Lee 2017; Lundberg et al. 2020). For the tree predictor $g(\mathbf{x})$, the SHAP value for feature k at $\mathbf{x}$ is

$$\phi_k(\mathbf{x}) = \sum_{S \subseteq \{1,\dots,p\} \setminus \{k\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \Big\{ \mathbb{E}\big[g(\mathbf{x}) \mid \mathbf{x}_S\big] - \mathbb{E}\big[g(\mathbf{x}) \mid \mathbf{x}_{S \cup \{k\}}\big] \Big\}, \tag{15}$$

which averages the marginal contribution of feature k over all subsets S of the remaining predictors. The term $\hat{\beta}_k(x_{ijk} - \bar{x}_k)$ is the centred linear contribution. Because the trees are fitted after adjusting for estimated random-effect contributions, the ranking describes predictor contributions to the fitted population component.

Computational cost.

Let $B(N, p, M)$ denote the computational cost of a single boosting fit with M trees. For the exact greedy algorithm, this cost is on the order of $MdN_{\text{nz}} \log N$, where N_{nz} is the number of non-zero entries in the design matrix (Chen and Guestrin 2016). Let $C_{\text{L}}(N, p)$ denote the cost of computing a lasso solution path. One outer iteration incurs a cost of $(V+1)\{B(N, p, \hat{M}) + C_{\text{L}}(N, p)\}$, in addition to $(V+1)\sum_K B(N, K, \hat{M}_K)$ associated with the screened candidates during the first two iterations, and an expectation-step cost of $O\{L_{\text{in}}(\sum_i n_i q^2 + nq^3)\}$. Training fold-specific models in geometrically increasing chunks mitigates the overhead of repeatedly copying the booster state. Moreover, when a screened model is ultimately selected, the resulting reduction in the effective predictor dimension yields a substantial decrease in computational burden, particularly in high-dimensional regimes (large p).

4 Theoretical properties

This section presents theoretical results that motivate the algorithm and help interpret the simulation findings. Proofs are in Appendix A, together with auxiliary results on the Newton step, the expectation step, the centring of the random effects, the restriction of linear predictions to the training range, and the cost of growing trees in chunks. Throughout, model (1) and Assumption 1 hold unless stated otherwise.

4.1 Alternating updates of the fixed and random effects

The first two results describe what happens when the expectation step uses the in-sample fit of a flexible learner. The first uses an idealised interpolating learner to demonstrate the mechanism; the second quantifies the retained random-effect signal for linear smoothers.

Proposition 1 (In-sample alternation with an interpolating learner) *Consider a random-intercept model ($q = 1$, $\mathbf{z}_{ij} = 1$, $n_i = m$) with $\sigma^2 > 0$ held fixed and $\boldsymbol{\beta}_z = 0$, and suppose the learner interpolates its training responses, $\hat{f}^{(t)}(\mathbf{x}_{ij}) = y_{ij} - \hat{b}_i^{(t-1)}$, while the random effects and their variance d_t are updated by (7) and (8) from the in-sample residuals $y_{ij} - \hat{f}^{(t)}(\mathbf{x}_{ij})$. Then $\hat{b}_i^{(t)} = k_t \hat{b}_i^{(t-1)}$ with $k_t = md_t/(md_t + \sigma^2) \in (0, 1)$, the sequence d_t is non-increasing for $t \geq 2$, and $d_t \to 0$ and $\hat{b}_i^{(t)} \to 0$ for every i.*

Proposition 2 (Random effect retained by in-sample and cross-fitted residuals) *Let $\hat{\mathbf{f}} = \mathbf{S}\tilde{\mathbf{y}}$ be a linear smoother whose weights depend only on the predictors, applied to $\tilde{\mathbf{y}} = \mathbf{y}$ with $\hat{\mathbf{b}} = \mathbf{0}$ in a random-intercept model, and let $\pi_{ij} = \sum_l S_{ij,il}$ be the total weight that the fitted value at (i, j) gives to measurements of subject i. Then b_i enters the residual $y_{ij} - \hat{f}(\mathbf{x}_{ij})$ with coefficient $1 - \pi_{ij}$. For a cross-fitted smoother, whose row (i, j) assigns zero weight to the fold containing (i, j), has entries summing to one, and, under exchangeable measurements, distributes its weight evenly in expectation across the fold model's training measurements, $\mathbb{E}\pi_{ij}^{\text{cf}} = n_i/N$, which equals $1/n$ in a balanced design.*

A highly flexible boosted ensemble can place substantial weight on y_{ij} when fitting $\hat{f}(\mathbf{x}_{ij})$. When π_{ij} is close to one, the in-sample residual retains little of b_i. Under the cross-fitted smoother assumptions, the retained fraction is $1 - O(1/n)$ in a balanced design. The next result describes what the expectation step then estimates.

Proposition 3 (Cross-fitted working model) *The cross-fitted residual satisfies $\mathbf{y}_i - \hat{f}_i^{\text{cf}} - \mathbf{Z}_i\boldsymbol{\beta}_z = \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i + \mathbf{e}_i$ with $\mathbf{e}_i = \mathbf{f}_i - \hat{f}_i^{\text{cf}}$, and, conditionally on the measurements outside $\mathcal{F}_v$, ε_{ij} is independent of e_{ij} for every $(i, j) \in \mathcal{F}_v$. If the e_{ij} have mean zero and variance τ^2 and are independent of the random effects and of each other, a converged solution of the EM updates (7)–(8) applied to these residuals is the maximum likelihood estimate of $(\mathbf{D}, \sigma^2 + \tau^2)$ in a linear mixed model; the out-of-fold error of the fixed effect is attributed to the residual variance and not to $\mathbf{D}$.*

Remark 1 (Monotonicity) With the in-sample expectation step and a fixed structure of f, the alternation is a generalised EM algorithm and the log-likelihood (3) does not decrease, provided the boosting step does not decrease the EM surrogate. The cross-fitted algorithm does not retain this monotonicity guarantee because $\hat{f}^{\mathrm{cf}}$ does not maximise the same surrogate. The best-state rule in Section 3.3 retains the fit with the smallest recorded out-of-fold error, rather than necessarily returning the final iteration.

Proposition 4 (Rate of the variance-component EM) *In a balanced random-intercept model with $n_i = m$, known residual variance s^2 and fixed offset, let $\bar{r}_i$ be the mean residual of subject i and $\overline{\bar{r}^2} = n^{-1}\sum_i \bar{r}_i^2$. The EM map for the intercept variance is $g(d) = d^2(d+a)^{-2}\,\overline{\bar{r}^2} + da/(d+a)$ with $a = s^2/m$, its interior fixed point is the maximum likelihood estimate $d^\star = \overline{\bar{r}^2} - a$, and $g'(d^\star) = 1-\rho^2$, where $\rho = d^\star/(d^\star + s^2/m)$ is the reliability of a subject mean.*

In the simulation design of Section 5, with training measurements per subject, $d^\star = 0.5$ and $s^2 = 0.25$, the convergence factor is approximately 0.07, so only a few inner iterations are needed. When the fixed-effect fit remains inaccurate, s^2 can be large and the factor approaches one. This motivates repeating the inner updates for each boosting fit, subject to the iteration limit and convergence tolerance.

4.2 Selection using out-of-fold loss

Theorem 1 (Oracle inequality for the out-of-fold choice) *For a fold v and a finite set $\mathcal{G}$ of candidate fixed-effect specifications. Let $\mathcal{D}^{(-v)}$ be the measurements outside $\mathcal{F}_v$, and suppose that the fold models and the offsets of the held-out measurements are functions of $\mathcal{D}^{(-v)}$, that the squared error of a held-out measurement lies in $[0, B]$, and that each subject contributes at most $m_{\max}$ measurements to $\mathcal{F}_v$. Let $\widehat{\mathcal{L}}_v(\mathfrak{f})$ be the out-of-fold loss of candidate $\mathfrak{f}$ on $\mathcal{F}_v$, $R_v(\mathfrak{f}) = \mathbb{E}\{\widehat{\mathcal{L}}_v(\mathfrak{f}) \mid \mathcal{D}^{(-v)}\}/|\mathcal{F}_v|$ its conditional risk and $\hat{\mathfrak{f}} = \arg\min_{\mathcal{G}} \widehat{\mathcal{L}}_v$. For every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$R_v(\hat{\mathfrak{f}}) \le \min_{\mathfrak{f}\in\mathcal{G}} R_v(\mathfrak{f}) + \frac{2Bm_{\max}n}{|\mathcal{F}_v|}\sqrt{\frac{\log(2|\mathcal{G}|/\delta)}{2n}}.$$

Because $|\mathcal{F}_v| \approx N/V$ and $N = n\bar{m}$ with $\bar{m}$ the mean number of measurements per subject, the remainder is approximately $(2BVm_{\max}/\bar{m})\{\log(2|\mathcal{G}|/\delta)/(2n)\}^{1/2}$, of order $\{\log|\mathcal{G}|/n\}^{1/2}$ in the number of subjects, which is the effective sample size of clustered data. The logarithmic dependence on $|\mathcal{G}|$ means that evaluating many candidate tree counts increases the theoretical remainder only slowly. In MEXGB the training offsets of the fold models come from the previous expectation step, which used all measurements of each subject; the theorem then holds conditionally on these offsets, with the influence of a single held-out measurement on a predicted random effect described as being of order $1/n_i$.

Proposition 5 (Unbiased comparison of screened candidates) *For each fold v, the predictor ranking used to build the screened fold models is a function of $\mathcal{D}^{(-v)}$ and of the offsets. Conditionally on $\mathcal{D}^{(-v)}$ the out-of-fold loss of every screened candidate is therefore an unbiased estimator of $|\mathcal{F}_v|$ times its conditional risk, and Theorem 1 applies to the choice among the screened sets and the set of all predictors.*

Proposition 6 (Largest spurious gain) *Consider a node of m measurements whose residuals are independent $\mathcal{N}(0, \sigma^2)$ and independent of r inactive predictors, with each predictor split at its median, with $h \equiv 1$ and $\lambda = \gamma = 0$. The gain (6) of the split on predictor k is $\frac{1}{2}\sigma^2 Z_k^2$ with $Z_k \sim \mathcal{N}(0,1)$, and*

$$\mathbb{E} \max_{k \le r} \mathrm{Gain}_k \le \tfrac{1}{2}\sigma^2 \big\{ 2 \log(2r) + 1 \big\}.$$

Example 1 With $r = 2000$ the bound is $8.8\sigma^2$, and after screening to $K = 20$ it is $4.2\sigma^2$; the greedy search over split points increases both values without changing the logarithmic dependence on r.

Proposition 7 (Out-of-fold error as a prediction risk) *Suppose that measurements from a subject are exchangeable conditional on the random effect. Condition on the fold models. For $(i,j) \in \mathcal{F}_v$, the squared error $\{y_{ij} - \hat{f}^{(-v)}(\mathbf{x}_{ij}) - \mathbf{z}_{ij}^\top \boldsymbol{\beta}_z - \mathbf{z}_{ij}^\top \hat{\mathbf{b}}_i^{(-v)}(c\mathbf{D})\}^2$ has the same conditional expectation as the squared error of the same predictor for a new measurement of subject i. Hence $\mathcal{E}(c)/N$ is an unbiased estimator of the average error of predicting new measurements of the training subjects.*

Proposition 8 (Optimality of the true covariance under the model) *If f, $\boldsymbol{\beta}_z$, $\mathbf{D}$ and σ^2 are the true values, then for every held-out measurement $\mathbb{E}\{y_{ij} - f(\mathbf{x}_{ij}) - \mathbf{z}_{ij}^\top \boldsymbol{\beta}_z - \mathbf{z}_{ij}^\top \hat{\mathbf{b}}_i^{(-v)}(c\mathbf{D})\}^2$ is minimised at $c = 1$, and so is $\mathbb{E}\,\mathcal{E}(c)$.*

4.3 Consistency

Theorem 2 (Fisher consistency of the fixed-effect update) *If the offsets are the true random-effect contributions $o_{ij} = \mathbf{z}_{ij}^\top(\boldsymbol{\beta}_z + \mathbf{b}_i)$, the population criterion $R(g) = \mathbb{E}\{y - o - g(\mathbf{x})\}^2$ is minimised over measurable functions g by $g = f$, uniquely up to sets of probability zero, and $R(g) - R(f) = \mathbb{E}\{g(\mathbf{x}) - f(\mathbf{x})\}^2$.*

Theorem 3 (Variance components with an estimated fixed effect) *Let $\boldsymbol{\beta}_z$ be known, $n_i \le m_{\max}$, and $\boldsymbol{\theta} = (\mathbf{D}, \sigma^2)$ range over a compact set Θ with $\sigma^2 \ge \sigma^2_{\min} > 0$, and let $\boldsymbol{\theta}_0$ be the unique maximiser over Θ of $\mathbb{E}\ell_i(\boldsymbol{\theta}; f)$, where*

$$\ell_i(\boldsymbol{\theta}; \tilde{f}) = -\tfrac{1}{2}\big\{ \log|\mathbf{V}_i| + (\mathbf{y}_i - \tilde{\mathbf{f}}_i - \mathbf{Z}_i\boldsymbol{\beta}_z)^\top \mathbf{V}_i^{-1} (\mathbf{y}_i - \tilde{\mathbf{f}}_i - \mathbf{Z}_i\boldsymbol{\beta}_z) \big\}.$$

If $\Delta_N = N^{-1}\sum_{i,j}\{\hat{f}(\mathbf{x}_{ij}) - f(\mathbf{x}_{ij})\}^2 \to 0$ in probability, every maximiser $\hat{\boldsymbol{\theta}}$ of $n^{-1}\sum_i \ell_i(\boldsymbol{\theta}; \hat{f})$ over Θ converges in probability to $\boldsymbol{\theta}_0$ as $n \to \infty$.

The theorem requires consistency of the fixed-effect estimator in empirical L_2, without specifying a convergence rate. Under its assumptions, it also applies to $\hat{f} = \hat{f}^{\mathrm{cf}}$. Consistency of boosting with early stopping is established by Zhang and Yu (2005), and for sparse linear and additive models in high dimensions by the results collected in Bühlmann and van de Geer (2011).

4.4 Comparison with competing estimators

The next result gives a lower bound for predictors that do not use subject identity, subject to the stated conditional mean assumption.

Theorem 4 (Error of predictors that ignore subject identity) *Let $(\mathbf{x}^*, \mathbf{z}^*, y^*)$ be a new measurement of a subject i seen in training, and let $\hat{g}$ be any predictor computed from $\mathbf{x}^*$, $\mathbf{z}^*$ and the training data without using subject identity, and suppose that $\mathbb{E}(\mathbf{z}^{*\top}\mathbf{b}_i + \varepsilon^* \mid \mathbf{x}^*, \mathbf{z}^*, \hat{g}) = 0$. Then*

$$\mathbb{E}(y^* - \hat{g})^2 \geq \sigma^2 + \mathbb{E}\big(\mathbf{z}^{*\top}\mathbf{D}\mathbf{z}^*\big), \tag{16}$$

The mixed-model predictor using the true f, $\boldsymbol{\beta}_z$, and variance components has mean squared error $\sigma^2 + \mathbb{E}(\mathbf{z}^{*\top}\mathbf{C}_i\mathbf{z}^*)$, with $\mathbf{C}_i = (\mathbf{D}^{-1} + \sigma^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i)^{-1} \preceq \mathbf{D}$ and $\mathbf{C}_i \to \mathbf{0}$ as the smallest eigenvalue of $\mathbf{Z}_i^\top\mathbf{Z}_i$ diverges.

Example 2 In the simulation design of Section 5, $\mathbf{z} = (1, t)^\top$ with t uniform on 20 equally spaced points of $[0, 1]$, $D_{00} = 0.5$, $D_{11} = 3$, $D_{01} = 0.6\sqrt{1.5} = 0.735$ and $\sigma^2 = 0.25$. Then $\mathbb{E}t = 0.5$, $\mathbb{E}t^2 = 0.342$ and $\mathbb{E}(\mathbf{z}^\top\mathbf{D}\mathbf{z}) = 0.5 + 2(0.735)(0.5) + 3(0.342) = 2.26$, giving an expected MSE lower bound of 2.51 for predictors satisfying the assumptions of Theorem 4.

The next result describes how fixed-effect estimation error affects prediction for a previously observed subject and identifies the component that random-effect prediction can correct.

Proposition 9 (Error decomposition for a previously observed subject) *Let the variance components and $\boldsymbol{\beta}_z$ equal their true values, let $\hat{f}$ be independent of the random effect and errors of subject i, and let $\hat{\mathbf{b}}_i = \mathbf{D}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}(\mathbf{y}_i - \hat{\mathbf{f}}_i - \mathbf{Z}_i\boldsymbol{\beta}_z)$. For a new measurement $(\mathbf{x}^*, \mathbf{z}^*, y^*)$ of subject i write $e(\mathbf{x}) = f(\mathbf{x}) - \hat{f}(\mathbf{x})$, $e^* = e(\mathbf{x}^*)$, $\mathbf{e}_i = (e(\mathbf{x}_{i1}), \ldots, e(\mathbf{x}_{in_i}))^\top$ and $\mathbf{k}^\top = \mathbf{z}^{*\top}\mathbf{D}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}$. Then*

$$\mathbb{E}\big\{y^* - \hat{f}(\mathbf{x}^*) - \mathbf{z}^{*\top}(\boldsymbol{\beta}_z + \hat{\mathbf{b}}_i)\big\}^2 = \sigma^2 + \mathbb{E}\big(\mathbf{z}^{*\top}\mathbf{C}_i\mathbf{z}^*\big) + \mathbb{E}\big(e^* - \mathbf{k}^\top\mathbf{e}_i\big)^2. \tag{17}$$

The fixed-effect error enters through $e^* - \mathbf{k}^\top\mathbf{e}_i$: the predicted random effect subtracts a weighted combination of the subject's own fixed-effect errors. An error that is constant within a subject or linear in time can be partly absorbed by the random intercept and slope. Errors that vary with measurement-specific predictors are less readily corrected. The accuracy of $\hat{f}$ at the level of individual measurements is therefore important when comparing mixed-effects methods with the same random-effect structure.

Corollary 1 (Error floor of a misspecified linear mixed model) *Suppose the predictors of different measurements are independent and identically distributed and independent of $(\mathbf{z}, \mathbf{b}_i)$, and let $\hat{f}$ be the L_2 projection $a^\circ + \mathbf{x}^\top\boldsymbol{\beta}^\circ$ of f on affine functions. Then the error in (17) is at least $\sigma^2 + \mathbb{E}(\mathbf{z}^{*\top}\mathbf{C}_i\mathbf{z}^*) + \Delta_{\mathrm{lin}}(f)$ with $\Delta_{\mathrm{lin}}(f) = \min_{a,\boldsymbol{\beta}} \mathbb{E}\{f(\mathbf{x}) - a - \mathbf{x}^\top\boldsymbol{\beta}\}^2$, which is positive unless f is affine.*

MEGB, MERF, REEMforest, GPBoost, and cluster-blind ensemble methods approximate f using functions that are piecewise constant over the cells of axis-aligned partitions. The following two results characterise the approximation error incurred when representing a linear signal under this modelling constraint, thereby providing theoretical motivation for incorporating an explicit linear component.

Theorem 5 (Piecewise-constant approximation of a linear function) *Let $\mathbf{x} = (\mathbf{x}_A, \mathbf{x}_B)$ with $\mathbf{x}_A = (x_1, \ldots, x_s)$ uniform on $[0,1]^s$ and independent of $\mathbf{x}_B$, and let $f(\mathbf{x}) = \sum_{k=1}^{s} \beta_k x_k$. Let g be a sum of M regression trees of depth at most d, and let K_k be the number of distinct split points on x_k in the ensemble. Then*

$$\mathbb{E}\{f(\mathbf{x}) - g(\mathbf{x})\}^2 \;\geq\; \sum_{k=1}^{s} \frac{\beta_k^2}{12(K_k+1)^2} \;\geq\; \frac{\bigl(\sum_{k=1}^{s} |\beta_k|^{2/3}\bigr)^3}{12\{M(2^d-1)+s\}^2}. \tag{18}$$

Increasing the number of trees can reduce approximation error, but the trees must also be estimated from data. For a partition fixed in advance, the trade-off between approximation and estimation error can be expressed explicitly.

Theorem 6 (Partition estimator of a linear signal) *Let $y = \beta x + \varepsilon$ with x uniform on $[0,1]$ and ε independent of x with mean zero and variance σ^2, and let $\hat{g}$ be a partition estimator based on $K \leq N$ equal-length intervals, taking the mean observed response within each interval. Conditionally on the event that every interval contains at least one of the N observations,*

$$\mathbb{E}\Bigl[\int_0^1 \{\hat{g}(x) - \beta x\}^2 dx \;\Big|\; x_1, \ldots, x_N\Bigr] \;\geq\; \frac{\beta^2}{12K^2} + \frac{\sigma^2 K}{N} \;\geq\; 3\Bigl(\frac{\beta^2\sigma^4}{48N^2}\Bigr)^{1/3},$$

whereas the least-squares line has integrated risk $2\sigma^2 N^{-1}\{1 + o(1)\}$.

Theorem 6 gives a lower bound of order $N^{-2/3}$ for the specified partition estimator, compared with the N^{-1} risk of a correctly specified linear fit. The estimation-risk result does not directly cover boosted trees, whose partitions are selected from the data. The approximation bound in Theorem 5, however, applies to ensembles with a given number of split points under its stated design assumptions. For a sparse linear model with s active coefficients, standard lasso theory gives prediction-error rates of order $s \log p / N$ under appropriate compatibility and noise conditions (Bickel et al. 2009; Bühlmann and van de Geer 2011). In MEXGB, the linear component is fitted first and the trees then estimate the remaining signal. If the signal is linear, the out-of-fold criterion can select $\hat{M} = 0$, leaving a relaxed lasso as the fixed-effect estimator.

4.5 Special cases

Corollary 2 (Special cases) *(i) With $\widehat{\mathbf{D}} = \mathbf{0}$ and $\boldsymbol{\beta}_z = \mathbf{0}$ held fixed and no linear component, MEXGB is XGBoost with the squared-error loss and a number of trees chosen by cross-validation. (ii) With trees only, $\lambda = 0$, first-order leaf weights and a fixed number of trees, the algorithm is a cross-fitted version of MEGB. (iii) With the linear component only ($M = 0$) and $\lambda_{\mathrm{L}} = 0$, the algorithm is the expectation–conditional maximisation algorithm of the linear mixed model.*

5 Simulation study

The simulation study has two parts. The first compares the predictive accuracy and variable-selection performance of MEXGB with those of ten competing methods across 13 scenarios. The second evaluates estimation of the random-effects structure and fixed effect and examines the contributions of the linear component and cross-fitted expectation step.

5.1 Data-generating process

The design extends the published MEGB and longitudinal random forest studies (Capitaine et al. 2021; Olaniran et al. 2025) by increasing the sample size. Each dataset has $n = 50$ subjects observed at 20 equally spaced times $t_{ij} \in \{0, 1\}$, so $N = 1000$. For each measurement, the p predictors are drawn independently of those at other measurements from $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ with $\Sigma_{kk'} = \rho^{|k-k'|}$. Six predictors are active, with either a linear or a nonlinear fixed effect,

$$f_{\text{lin}}(\mathbf{x}) = 2x_1 - 2x_2 + 1.5x_3 - 1.5x_4 + x_5 - x_6,$$
$$f_{\text{non}}(\mathbf{x}) = 2\sin(\pi x_1 x_2) + 4(x_3 - 0.5)^2 + 2x_4 + e^{x_5} + x_6 x_1.$$

The nonlinear function combines a sinusoidal interaction, a quadratic term, a linear term, an exponential term, and a product interaction. It is therefore not representable by a purely linear fixed-effect specification. The responses are

$$y_{ij} = f(\mathbf{x}_{ij}) + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij}, \qquad (b_{0i}, b_{1i})^\top \sim \mathcal{N}_2(\mathbf{0}, \mathbf{D}), \qquad \text{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma^2\rho^{|j-j'|},$$

with $D_{00} = 0.5$, $D_{11} = 3$ and correlation 0.6 between intercept and slope. The within-subject errors follow a first-order autoregressive covariance structure. The working model of every mixed-effects method assumes independent errors given the random effects, so the serial correlation is a misspecification shared by all of them.

In the primary simulation study, the two fixed-effect specifications are evaluated in a full-factorial design with $p \in \{6, 170, 2000\}$, holding $\rho = 0.6$ and $\sigma = 0.5$ constant. A complementary sensitivity analysis, restricted to the nonlinear fixed-effect specification with $p = 170$, varies $\rho \in \{0, 0.2, 0.6, 0.8\}$, a parameter governing both the inter-predictor correlation and the serial correlation of the error process and considers $\sigma \in \{0.5, 1.5\}$. For $\rho = 0$, within-subject dependence arises exclusively through the random effects. Combining the primary and sensitivity analyses yields 13 distinct experimental scenarios, as one sensitivity configuration coincides with a primary-study configuration.

For each scenario, each simulated dataset was subjected to 100 repeated random 70/30 train–test splits. Within each split, 30% of the measurements per subject (6 of 20) were allocated to the test set. As a result, both the training set (700 measurements) and the test set (300 measurements) included observations from all subjects. For each split, the mixed-effects model was fit using the training data only, and predictions for the test observations were generated using the estimated fixed effects together with the subject-specific random effects estimated from the training data.

5.2 Competing methods

Table 1 summarises the competing methods and their hyperparameter configurations. To improve comparability among boosting-based competitors, we imposed a common boosting budget of 500 trees, a learning rate of 0.05, a tree depth of 2, a minimum of two observations per terminal node, and no subsampling. This budget was selected in a preliminary tuning exercise over the number of trees (200 – 500), the depth of the tree (2 – 8) and the learning rate (0.01 – 0.2 for MEXGB, MEGB, GBM and GPBoost), and the resulting values were then fixed for all data splits and simulation scenarios. Consequently, the comparison is not driven by differences in the nominal boosting budget, although methods still differ in implementation and estimation strategy.

MEXGB and MEGB are both fitted with cross-fitting ($V = 5$), using the E-step described in Section 3.3 with LMM-based initialisation; MEGB uses 50 inner EM steps. The comparison therefore primarily evaluates differences in the fixed-effect base learners and in the out-of-fold model-selection mechanisms, with cross-fitting enabled for both approaches. MERF and REEMforest are fitted via `LongituRF` (Capitaine et al. 2021), excluding the stochastic-process component, and use the same random-intercept and random-slope structure as MEXGB and MEGB.

The remaining method-specific hyperparameters were tuned by 10-fold repeated cross-validation and then fixed across data splits. For GPBoost, we additionally optimised the Gaussian-process kernel hyperparameters (Matérn length scale: 0.1–10). For glmmLasso, the regularisation parameter $\lambda_{\text{glmmLasso}}$ was selected in the range 10^{-4} to 10^{2}. For XGBoost, the regularisation parameter was fixed at $\lambda_{\text{XGB}} = 1$. For RF, MERF, and REEMforest, `mtry` was fixed at $\lfloor p/3 \rfloor$. For LMM, the variance components were estimated using restricted maximum likelihood (REML).

Table 1 Methods compared in the simulation study. "Common budget" denotes 500 trees, learning rate 0.05, tree depth 2, at least two observations per terminal node, and no subsampling; these values were held fixed across all splits and scenarios. "Int. + slope" denotes a random intercept and random slope.

Method	Settings	Random effects
MEXGB	`xgboost`, common budget, relaxed lasso, cross-fitting $V = 5$	int. + slope
MEGB	`gbm`, common budget, cross-fitted EM ($V = 5$, 50 inner steps)	int. + slope
MERF, REEMforest	500 trees, `mtry` $= \lfloor p/3 \rfloor$	int. + slope
GPBoost	LightGBM, common budget, Matérn kernel	int. + slope
XGBoost	common budget, $\lambda_{\text{XGB}} = 1$	none
GBM	`gbm`, common budget	none
RF	500 trees, `mtry` $= \lfloor p/3 \rfloor$	none
mboost	`glmboost`, common budget	none
glmmLasso	$\lambda_{\text{glmmLasso}}$ by CV	int. + slope
LMM	`nlme`, REML	int. + slope

5.3 Metrics

Prediction accuracy is measured by the test mean squared error, MSE $= N_{\text{test}}^{-1} \sum_{(i,j)\in\text{test}} (y_{ij} - \hat{y}_{ij})^2$, averaged over the 100 splits. Because all methods are evaluated on the same splits, they are also compared in pairs: for each split we compute the difference between the MSE of MEXGB and that of a competitor, and report its mean, the percentage of splits in which MEXGB is more accurate, and the p-value of the two-sided Wilcoxon signed-rank test. The 100 splits of a scenario come from 10 datasets and are therefore not independent; we also compare the mean MSE over the ten splits of each dataset, giving ten independent dataset-level paired comparisons per scenario. Variable selection is assessed using the top six predictors with positive importance. The true positive rate (TPR) is the percentage of the six active predictors selected, and the false positive rate (FPR) is the percentage of the $p-6$ inactive predictors selected. For MEXGB the ranking uses (14); for the other methods it uses their standard importance measures (relative influence, permutation importance, absolute coefficients or t-statistics).

The estimation study of Section 5.4.4 uses the full dataset of each replicate and assesses estimation of the random-effects structure using

$$\text{mse}_{\mathbf{D}} = \|\widehat{\mathbf{D}} - \mathbf{D}\|_F^2, \qquad \text{mse}_{\mathbf{b}} = \frac{1}{n}\sum_{i=1}^{n} \|\hat{\mathbf{b}}_i - \mathbf{b}_i\|^2,$$

where $\mathbf{b}_i$ are the realised random effects. Population-mean estimation is assessed using $\text{mse}_f = G^{-1}\sum_{g=1}^{G}\{\hat{f}(\mathbf{x}_g) + \mathbf{z}_g^\top \hat{\boldsymbol{\beta}}_z - f(\mathbf{x}_g)\}^2$ evaluated at $G = 5000$ newly generated predictor vectors and measurement times.

5.4 Results

5.4.1 Linear scenarios

Table 2 reports test MSE and selection rates for the linear scenarios. Figure 1 shows the MSE distributions across splits, and Table 3 presents paired comparisons.

MEXGB attains the smallest mean test MSE at each predictor dimension, namely 0.213, 0.219, and 0.229 for $p = 6$, 170, and 2000, respectively, and it outperforms the strongest competing method on all 100 data splits. The best-performing competitors are the linear mixed model for $p = 6$ (MSE 0.528), GPBoost for $p = 170$ (MSE 0.945), and MEGB for $p = 2000$ (MSE 1.342); relative to these baselines, MEXGB reduces MSE by 60%, 77%, and 83%, respectively. Three main empirical patterns emerge.

First, MEXGB exhibits only a mild increase in prediction error as p grows, whereas the error of tree-only approaches increases substantially. In particular, the tree-only mixed models MEGB, MERF, and REEMforest incur errors that are approximately four- to fifteen-fold larger than those of MEXGB. This behavior is consistent with the approximation–estimation trade-offs described in Theorems 5 and 6: tree ensembles approximate linear effects via step functions, while an explicit linear component enables direct coefficient estimation. In MEXGB, the relaxed lasso directly estimates the six linear coefficients, and the out-of-fold selection procedure typically yields solutions with few or no trees.

Table 2 Linear scenarios: test mean squared error (MSE; mean and standard error over 100 train/test splits), and true and false positive rates (TPR, FPR, in %) of the six top-ranked predictors. The lowest MSE in each column is in bold; – indicates that the method fitting was not available (p too large). For $p = 6$ all six predictors are active, so no inactive predictor can be selected and the FPR is 0 for every method.

	$p = 6$			$p = 170$			$p = 2000$		
Method	MSE	TPR	FPR	MSE	TPR	FPR	MSE	TPR	FPR
MEXGB	**0.213 (0.002)**	100.0	0.00	**0.219 (0.002)**	100.0	0.00	**0.229 (0.002)**	100.0	0.00
MEGB	0.842 (0.008)	100.0	0.00	0.989 (0.012)	100.0	0.00	1.342 (0.017)	100.0	0.00
GPBoost	0.639 (0.010)	100.0	0.00	0.945 (0.011)	89.2	0.40	1.473 (0.020)	28.8	0.21
MERF	1.730 (0.015)	100.0	0.00	3.019 (0.028)	85.3	0.54	3.523 (0.027)	78.5	0.06
REEMforest	1.716 (0.015)	100.0	0.00	3.024 (0.027)	85.7	0.52	3.515 (0.026)	79.0	0.06
XGBoost	3.040 (0.047)	100.0	0.00	3.686 (0.046)	98.7	0.05	4.271 (0.056)	85.7	0.04
GBM	3.078 (0.045)	100.0	0.00	3.675 (0.046)	98.7	0.05	4.244 (0.054)	85.0	0.05
RF	3.711 (0.047)	100.0	0.00	4.992 (0.048)	83.8	0.59	5.460 (0.045)	73.3	0.08
mboost	2.560 (0.041)	100.0	0.00	2.633 (0.033)	100.0	0.00	2.855 (0.046)	100.0	0.00
glmmLasso	1.402 (0.049)	84.5	0.00	4.085 (0.127)	92.0	0.29	–	–	–
LMM	0.528 (0.007)	100.0	0.00	–	–	–	–	–	–

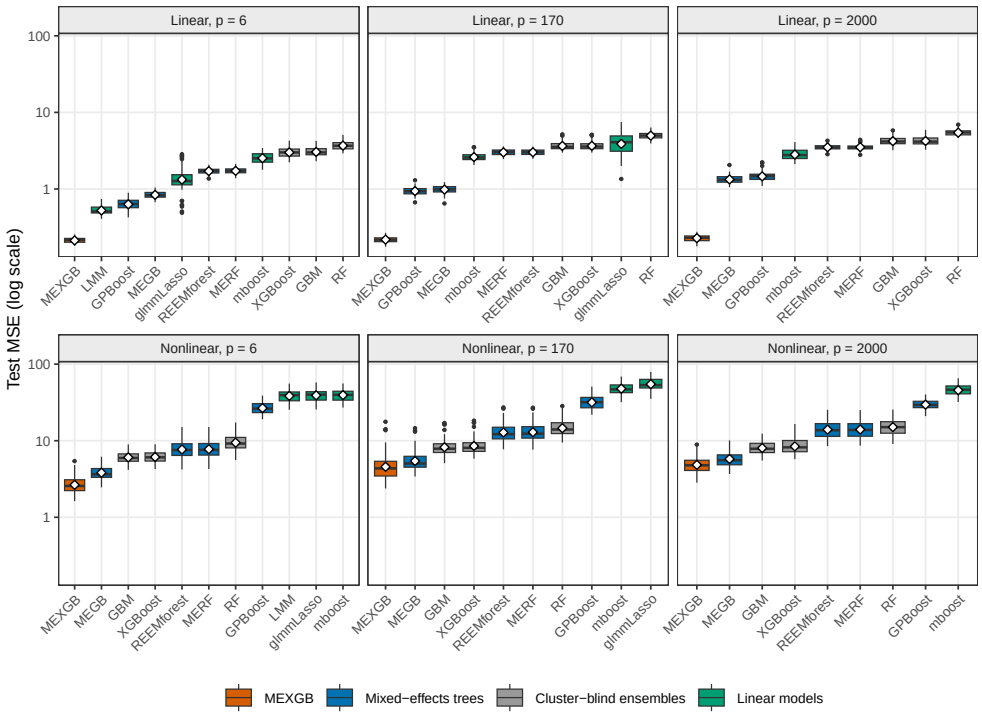


Fig. 1 Test MSE (logarithmic scale) over the 100 train/test splits of each main scenario. Within each panel the methods are ordered by mean MSE (diamonds), with the most accurate method on the left; colours indicate method families. Methods that were not fitted in a scenario are omitted.

Second, the linear benchmark methods are less accurate than MEXGB despite the fixed-effect mean being correctly specified. In addition, coefficient shrinkage may

Table 3 Paired comparison of MEXGB with the best competitor (by mean MSE), with MEGB and with XGBoost on identical train/test splits (main scenarios, 100 splits each). Diff.: mean of the MSE of MEXGB minus that of the competitor; Rel.: relative change of the mean MSE (%); Better: percentage of splits on which MEXGB has the lower MSE; p: two-sided Wilcoxon signed-rank test.

Scenario	Competitor	Diff.	Rel.	Better	p
Linear, $p = 6$	LMM	-0.314	-60	100	< 0.001
	MEGB	-0.628	-75	100	< 0.001
	XGBoost	-2.827	-93	100	< 0.001
Linear, $p = 170$	GPBoost	-0.726	-77	100	< 0.001
	MEGB	-0.770	-78	100	< 0.001
	XGBoost	-3.466	-94	100	< 0.001
Linear, $p = 2000$	MEGB	-1.113	-83	100	< 0.001
	XGBoost	-4.042	-95	100	< 0.001
Nonlinear, $p = 6$	MEGB	-1.15	-30	100	< 0.001
	XGBoost	-3.47	-56	100	< 0.001
Nonlinear, $p = 170$	MEGB	-0.77	-13	85	< 0.001
	XGBoost	-3.88	-44	100	< 0.001
Nonlinear, $p = 2000$	MEGB	-0.93	-16	93	< 0.001
	XGBoost	-3.63	-42	100	< 0.001

degrade performance for glmmLasso and mboost, and mboost does not incorporate random effects.

Third, cluster-blind ensembles yield MSEs between 3.0 and 5.5, exceeding the reference value of 2.51 reported in Example 2. Relative to XGBoost, MEXGB reduces MSE by 93–95%. Notably, MEXGB achieves an MSE below the marginal error variance $\sigma^2 = 0.25$. This can occur because within-subject autocorrelation renders portions of the test error predictable from neighboring training observations; correspondingly, the estimated random intercept and slope can absorb part of this shared variation.

Across all splits and all dimensions, the six truly active predictors are consistently ranked among the top six predictors by MEXGB, with no false positives. MEGB and mboost exhibit identical selection performance, whereas XGBoost and GBM recover only 85–86% of the active predictors at $p = 2000$, and GPBoost recovers only 29%.

5.4.2 Nonlinear scenarios

Table 4 reports prediction and variable-selection results for the nonlinear scenarios.

Across all three scenarios, MEXGB attains the lowest mean squared error (MSE), with values of 2.75, 4.93, and 5.00, followed by MEGB (3.90, 5.69, and 5.93). Split-level paired Wilcoxon tests indicate statistically significant differences in each scenario ($p < 0.001$). MEXGB yields lower error than MEGB on 100%, 85%, and 93% of the splits, corresponding to relative MSE reductions of 30%, 13%, and 16%, respectively. Two design features of MEXGB likely contribute to these gains: (i) the inclusion of a linear component, which can represent linear projections of other model terms, and (ii) selection of the number of trees via out-of-fold loss rather than fixing this hyperparameter a priori. Section 5.4.4 evaluates these potential mechanisms in the estimation study. The cluster-blind boosting methods (XGBoost: 6.2–8.8; GBM: 6.1–8.4) are more accurate than the mixed-effects forests (8.0–14.4), and MEXGB reduces the MSE of XGBoost by 42–56%. Linear methods exhibit substantially larger MSEs (39–56), consistent with the approximation error induced by a misspecified linear mean

Table 4 Nonlinear scenarios: test mean squared error (MSE; mean and standard error over 100 train/test splits), and true and false positive rates (TPR, FPR, in %) of the six top-ranked predictors. The lowest MSE in each column is in bold; – indicates that the method fitting was not available (p too large). For $p = 6$ all six predictors are active, so no inactive predictor can be selected and the FPR is 0 for every method.

	$p = 6$			$p = 170$			$p = 2000$		
Method	MSE	TPR	FPR	MSE	TPR	FPR	MSE	TPR	FPR
MEXGB	**2.75 (0.08)**	100.0	0.00	**4.93 (0.24)**	80.0	0.49	**5.00 (0.13)**	61.5	0.08
MEGB	3.90 (0.08)	100.0	0.00	5.69 (0.20)	80.8	0.70	5.93 (0.15)	64.3	0.11
GPBoost	27.07 (0.48)	50.0	0.00	32.26 (0.62)	47.3	0.18	29.90 (0.44)	47.7	0.07
MERF	8.03 (0.23)	100.0	0.00	13.46 (0.41)	87.3	0.46	14.42 (0.36)	80.0	0.06
REEMforest	7.97 (0.23)	100.0	0.00	13.34 (0.41)	87.5	0.46	14.41 (0.37)	79.3	0.06
XGBoost	6.21 (0.11)	100.0	0.00	8.81 (0.24)	75.5	0.90	8.62 (0.19)	63.7	0.11
GBM	6.13 (0.10)	100.0	0.00	8.44 (0.22)	76.7	0.85	8.16 (0.16)	61.5	0.12
RF	9.67 (0.23)	100.0	0.00	15.03 (0.39)	86.3	0.50	15.45 (0.35)	79.0	0.06
mboost	40.06 (0.74)	97.3	0.00	48.58 (0.90)	51.5	1.77	46.24 (0.75)	49.0	0.15
glmmLasso	39.74 (0.72)	86.8	0.00	55.66 (1.01)	70.7	1.07	–	–	–
LMM	39.13 (0.73)	100.0	0.00	–	–	–	–	–	–

as characterized in Corollary 1. GPBoost achieves MSEs of 27–32, which remain larger than those of the cluster-blind boosters under the same computational budget.

Identifying the active predictors is more challenging than in the linear settings. MEXGB recovers all six active predictors at $p = 6$, and achieves recovery rates of 80% at $p = 170$ and 62% at $p = 2000$, comparable to MEGB (81% and 64%) and to the cluster-blind boosters, whereas the forest-based approaches recover 79–88%. The underlying nonlinear signal includes interaction effects, which may partly explain the higher recovery rates observed for forests with deeper trees relative to the depth-2 boosting ensembles. Across methods, false positive rates remain below 2%.

5.4.3 Sensitivity to correlation and noise

Table 5 and Figure 2 show how prediction error changes with ρ and σ in the nonlinear scenario with $p = 170$.

Table 5 Sensitivity study (nonlinear fixed effect, $p = 170$): mean test MSE over 100 splits by the correlation parameter ρ (correlation of the predictors and serial correlation of the errors) and the error standard deviation σ. Last row: percentage of splits on which MEXGB has a lower MSE than MEGB. The lowest MSE in each column is shown in bold.

	$\sigma = 0.5$				$\sigma = 1.5$			
Method	$\rho = 0$	0.2	0.6	0.8	$\rho = 0$	0.2	0.6	0.8
MEXGB	**4.91**	**5.01**	**4.93**	**4.78**	**7.45**	**7.67**	**6.89**	**5.68**
MEGB	5.55	5.87	5.69	5.88	8.26	8.45	7.69	6.77
GPBoost	32.30	35.97	32.26	34.10	36.89	35.57	32.51	31.25
MERF	13.68	15.23	13.46	11.96	16.36	16.72	14.87	11.99
REEMforest	13.64	15.18	13.34	11.91	16.33	16.70	14.86	11.99
XGBoost	8.10	8.35	8.81	8.52	10.55	10.77	10.83	10.51
GBM	7.80	7.94	8.44	8.15	10.14	10.30	10.46	10.28
RF	14.97	16.16	15.03	13.13	17.30	17.60	16.69	14.41
mboost	39.24	42.89	48.58	56.34	43.13	41.69	46.94	53.02
glmmLasso	45.39	49.64	55.66	61.19	49.52	48.16	53.37	57.64
MEXGB < MEGB (%)	92	88	85	91	89	87	86	96

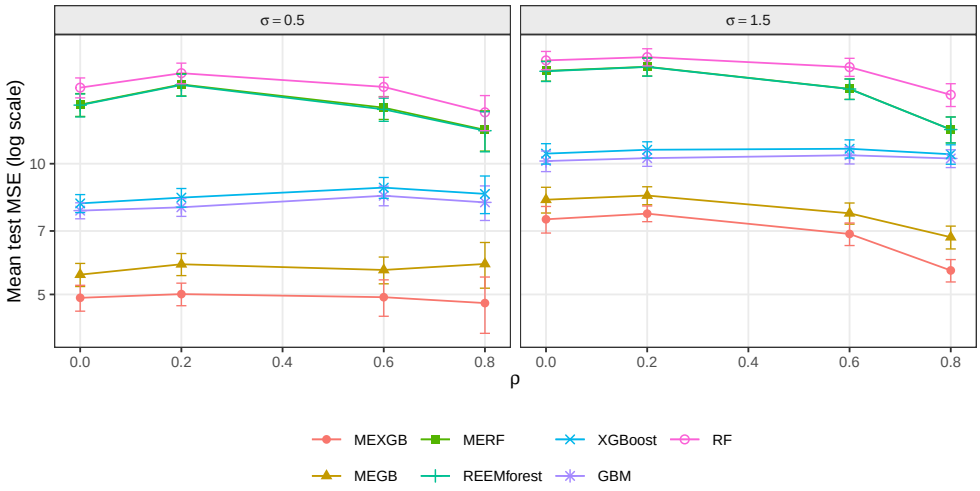


Fig. 2 Sensitivity study (nonlinear fixed effect, $p = 170$): mean test MSE (± 2 standard errors, logarithmic scale) by the correlation parameter ρ and the error standard deviation σ, for the tree-based methods.

Across all eight sensitivity settings, MEXGB exhibits the lowest prediction error and outperforms MEGB on 85–96% of the data splits. The mean paired MSE reductions ($MEXGB - MEGB$) range from 0.64 to 1.10 units, and split-wise Wilcoxon tests yield $p < 0.001$ in every setting. The relative performance ranking is stable across conditions: MEXGB $<$ MEGB $<$ GBM $\leq$ XGBoost.

Increasing σ from 0.5 to 1.5 increases the MSE of XGBoost by approximately 2.0–2.5 units when $\rho \leq 0.6$, which is close to the increment in the error variance, $1.5^2 - 0.5^2 = 2$. In contrast, when $\rho = 0.8$, the same increase in σ raises the MSE of MEXGB by only about 0.9 units, suggesting that strongly autocorrelated residuals partially emulate a subject-level effect that can be absorbed by the predicted random effects. When $\rho = 0$ (i.e., measurements are correlated only via the random effects), MEXGB improves upon XGBoost by 3.2 MSE units, exceeding the subject-level variance (2.26) in Example 2. This magnitude is consistent with benefits from both random-effect adjustment and improved estimation of the fixed-effect component.

5.4.4 Estimation accuracy and contributions of the MEXGB components

The second component of the study applies each method to the full dataset across 100 replicates under each of four simulation scenarios, defined by linear versus nonlinear fixed effects, $p \in \{6, 170\}$, $\rho = 0.6$, and $\sigma = 0.5$. Performance is evaluated by comparing the resulting estimates to the true variance components, the realised random effects, and the true fixed-effect function. In addition to MEXGB, MEGB, and a linear mixed-effects model, we consider two ablated variants of MEXGB: (i) *trees only*, which omits the linear component, and (ii) *in-sample E-step*, in which the expectation step is

computed using an in-sample booster fit with 500 trees (learning rate 0.05) rather than the cross-fitted booster. Table 6 summarises the median estimates in replicates and Figure 3 displays the empirical distributions of the estimated covariance parameters.

Table 6 Estimation of the random-effects structure and population mean: medians over 100 replicates of the estimated variance components, of $\text{mse}_{\mathbf{D}}$, $\text{mse}_{\mathbf{b}}$ and mse_f, and computation time in seconds. True values: $D_{00} = 0.5$, $D_{11} = 3$, $D_{01} = 0.73$; the errors have marginal variance 0.25 and are autocorrelated. Reduced versions of MEXGB: without the linear component (trees only), and with the expectation step computed from the in-sample fit of a 500-tree booster (in-sample E-step).

Scenario	Method	$\hat{D}_{00}$	$\hat{D}_{11}$	$\hat{D}_{01}$	$\hat{\sigma}^2$	$\text{mse}_{\mathbf{D}}$	$\text{mse}_{\mathbf{b}}$	mse_f	Time
Linear, $p = 6$	MEXGB	0.54	3.32	0.48	0.18	0.56	0.529	0.035	13
	MEXGB (trees only)	0.53	3.27	0.54	0.40	0.42	0.534	0.238	145
	MEXGB (in-sample E-step)	0.55	3.26	0.51	0.11	0.47	0.580	0.267	7
	MEGB	0.53	3.34	0.50	0.70	0.51	0.508	0.542	8
	LMM	0.56	3.39	0.49	0.18	0.60	0.532	0.032	0
Linear, $p = 170$	MEXGB	0.53	3.27	0.50	0.19	0.63	0.422	0.013	188
	MEXGB (trees only)	0.50	3.41	0.50	0.47	0.68	0.471	0.269	287
	MEXGB (in-sample E-step)	0.46	3.17	0.45	0.06	0.70	0.553	0.430	58
	MEGB	0.46	3.43	0.47	0.77	1.01	0.542	0.541	50
Nonlinear, $p = 6$	MEXGB	0.50	3.36	0.47	2.18	0.34	0.834	2.195	152
	MEXGB (trees only)	0.43	3.65	0.49	2.29	0.57	0.836	2.363	154
	MEXGB (in-sample E-step)	0.46	2.80	0.47	0.69	0.44	0.738	2.772	5
	MEGB	0.56	3.57	0.31	3.48	0.77	0.864	3.416	7
	LMM	0.37	4.99	0.00	38.33	23.71	2.689	40.652	0
Nonlinear, $p = 170$	MEXGB	0.63	2.93	0.93	3.12	1.57	0.803	2.792	230
	MEXGB (trees only)	0.54	2.97	0.95	3.36	0.93	0.811	3.047	182
	MEXGB (in-sample E-step)	0.39	2.46	0.66	0.47	0.99	0.983	4.133	60
	MEGB	0.53	3.48	0.85	4.18	1.70	0.875	3.991	51

Fixed effect. Incorporating a linear component markedly improves fixed-effect estimation when the underlying signal is linear. Specifically, for MEXGB, mse_f equals 0.035 and 0.013 at $p = 6$ and $p = 170$, respectively, compared with 0.24 and 0.27 for the trees-only variant and 0.54 for MEGB. At $p = 6$, the resulting accuracy is comparable to that of the correctly specified LMM (0.032). Across settings, MEXGB attains lower mse_f than its trees-only counterpart. Under nonlinear data-generating mechanisms, the two MEXGB variants perform similarly (2.20 versus 2.36, and 2.79 versus 3.05, respectively), and both substantially outperform MEGB (3.42, 3.99). These results indicate that, in nonlinear scenarios, the performance gains relative to MEGB are primarily driven by components common to both the full and trees-only MEXGB specifications.

Expectation step. Employing an in-sample expectation step enables the booster to absorb a larger share of within-subject variability. Consequently, the estimated residual variance decreases to 0.06–0.11 in the linear scenarios, compared with 0.18–0.19 for both MEXGB and the LMM, and to 0.47–0.69 in the nonlinear scenarios. This modification is accompanied by reduced fixed-effect estimation accuracy: mse_f increases to 0.27–0.43 under linear signals, and to 2.77 and 4.13 under nonlinear signals. With 500 trees of depth 2, the booster does not interpolate the training data; thus, the random-effect variances may be decreased without degenerating to zero (median $\hat{D}_{11}$ 2.46 versus 2.93 for MEXGB in the nonlinear scenario with $p = 170$). The impact on

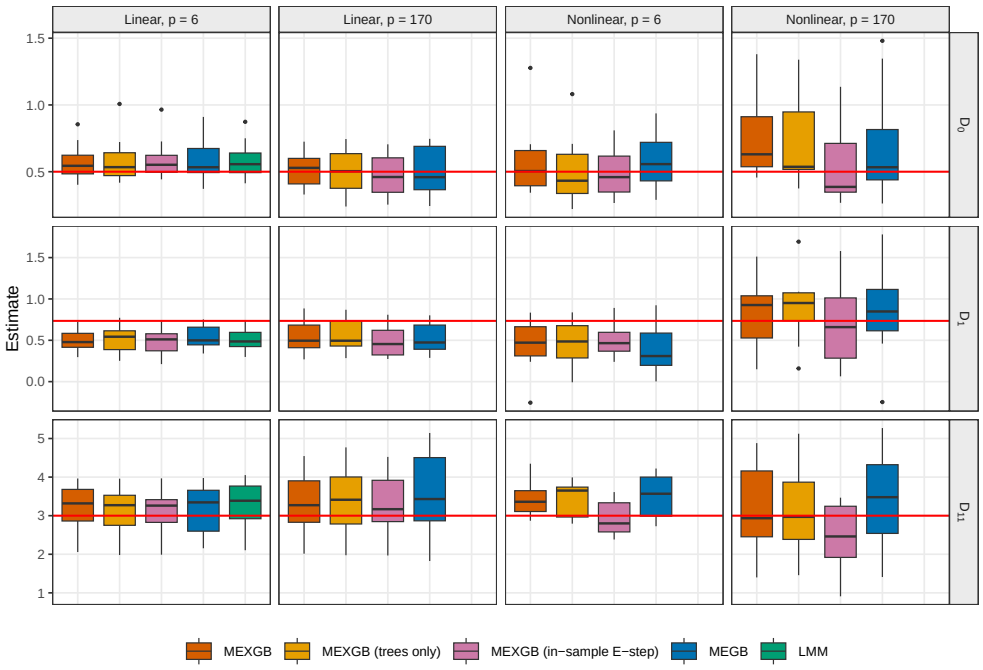


Fig. 3 Estimated random-effect variances and covariance over 10 replicates per scenario. Red lines indicate the true values ($D_{00} = 0.5$, $D_{01} = 0.73$, $D_{11} = 3$). The linear mixed model is shown only for the linear scenario with $p = 6$, in which its mean structure is correctly specified.

random-effect prediction is scenario-dependent: $\text{mse}_{\mathbf{b}}$ exceeds that of MEXGB in three scenarios (0.553 versus 0.422, and 0.983 versus 0.803 at $p = 170$), but is lower in the nonlinear setting with $p = 6$ (0.738 versus 0.834). Moreover, the residual-variance estimate based on cross-fitted predictions exceeds σ^2 when the signal is nonlinear (2.2 and 3.1), because it incorporates the out-of-fold fixed-effect prediction error, as stated in Proposition 3. MEGB, which is also cross-fitted, exhibits the same pattern (3.5, 4.2).

Random-effects structure. For prediction of the random effects, MEXGB attains the smallest $\text{mse}_{\mathbf{b}}$ at $p = 170$ (0.422, 0.803). At $p = 6$, MEGB (0.508 vs. 0.529 for the linear model) and the in-sample variant (0.738 vs. 0.834 for the nonlinear model) provide slightly improved random-effect predictions. Regarding estimation of the random-effects covariance matrix, MEXGB achieves the smallest $\text{mse}_{\mathbf{D}}$ in the linear setting with $p = 170$ and in the nonlinear setting with $p = 6$. In the linear setting with $p = 6$, all methods perform similarly (0.42–0.60). In contrast, in the nonlinear setting with $p = 170$, MEXGB overestimates D_{00} and D_{01} in a subset of replicates, leading to $\text{mse}_{\mathbf{D}} = 1.57$ compared with 0.93 for the trees-only variant and 1.70 for MEGB.

Under mean-model misspecification, the linear mixed model assigns part of the nonlinear signal to the random slope (yielding $\hat{D}_{11} = 4.99$), and its random effects are

estimated with low accuracy ($\text{mse}_{\mathbf{b}} = 2.69$). This demonstrates that fixed-effect misspecification can substantially degrade inference even when the random-effects design is correctly specified.

5.4.5 Computation and convergence

On a single CPU core, MEXGB requires approximately 6 s and 66 s for $p = 6$ (linear and nonlinear settings, respectively), about 2 minutes for $p = 170$, and 9–15 minutes for $p = 2000$ (Table C1 in Appendix C). At $p = 2000$, its runtime is of the same order of magnitude as that of MEGB, MERF, and REEMforest, but it is markedly slower than cluster-blind boosting approaches trained under a fixed computational budget. The dominant computational cost arises from fitting the $V+1$ boosting models within each outer iteration and from evaluating screened candidate models, particularly during the first two outer iterations.

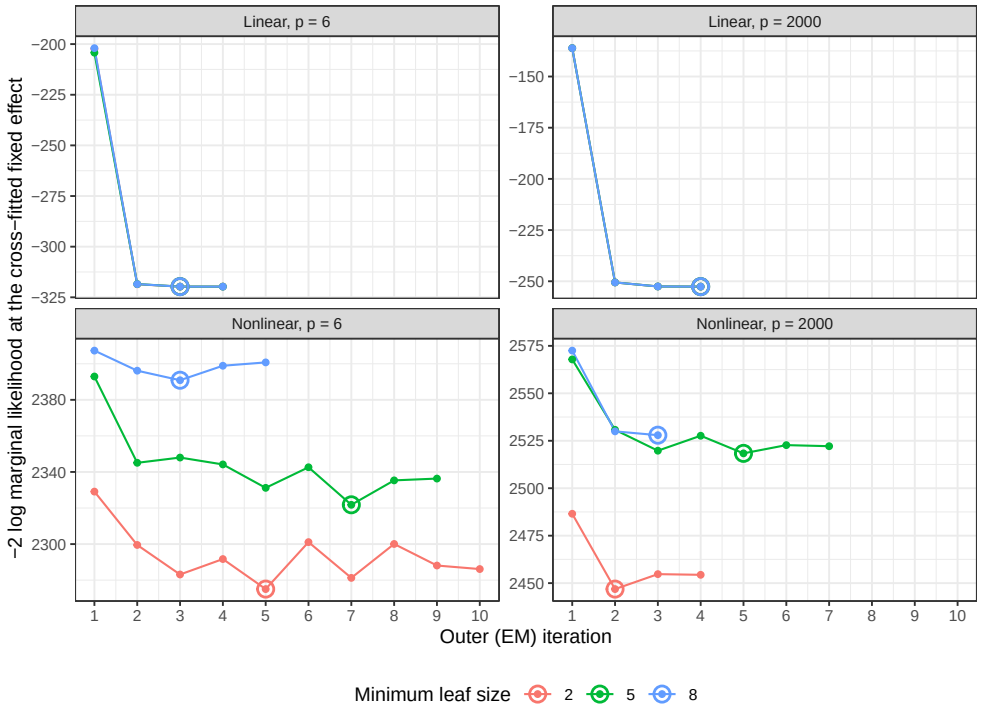


Fig. 4 Marginal -2 log-likelihood at the cross-fitted fixed effect over the outer iterations of MEXGB, for one dataset per scenario and minimum leaf sizes 2, 5 and 8 (2 is used in the study). Open circles indicate the iteration selected by the out-of-fold criterion.

Figure 4 reports the marginal -2 log-likelihood (3) evaluated at the cross-fitted fixed-effect estimate across outer iterations, shown for one representative dataset per scenario and three choices of the minimum leaf size. The criterion decreases steeply

during the first two iterations, coinciding with the transition of the random effects from their initial zero values to their predicted estimates, and subsequently stabilises or exhibits mild fluctuations, consistent with the behaviour expected under cross-fitting (Remark 1). In these examples, the out-of-fold criterion selects an iteration following the initial reduction, and termination occurs within ten outer iterations.

6 Applications

6.1 Data

We illustrate the method using three longitudinal datasets with continuous outcomes: two established low-dimensional growth datasets and one high-dimensional microbiome cohort.

BodyWeight (Crowder and Hand 1990), distributed with `nlme` (Pinheiro and Bates 2000), contains body weights (g) measured in 16 rats assigned to three diets (8, 4, and 4 rats, respectively). Each rat was measured 11 times between days 1 and 64 ($N = 176$). The predictors are time and two diet indicators ($p = 3$). The differences between rats are large relative to the measurement noise.

Soybean (Davidian and Giltinan 1995), also from `nlme`, contains average leaf weights per plant from 48 plots representing two genotypes planted in 1988, 1989, and 1990. Each plot was sampled 8 to 10 times between 14 and 84 days after planting ($N = 412$). The predictors are time, genotype and two year indicators ($p = 4$). Leaf weight is strongly right-skewed (mean 6.36 g, standard deviation 6.91 g, range 0.03–30.27 g) and is analysed on the log scale.

Sprockett is the gut-microbiome cohort of Tsimane horticulturalists of the Bolivian Amazon studied by Sprockett et al. (2020). We use stool samples from 193 individuals with at least three samples ($N = 1815$; 3 to 24 samples per individual, median 7). The predictors are 1879 amplicon sequence variants (ASVs), obtained by transforming the relative abundances to $\log(1 + x)$, centring the transformed abundances within each sample and removing the 10% of ASVs with the smallest variance. The response is the Shannon diversity of the sample (mean 2.19, standard deviation 0.89), and time is measured relative to each individual's first sample.

For mixed models (LMM, glmmlasso, MEXGB, MEGB, MERF, REEMforest, and GPBoost), the random-effects specification includes a subject-specific intercept and slope for time. Each dataset is split 100 times at random (ten repetitions of ten splits): in every split 30% of the measurements of each subject, and at least one, form the test set, with at least one measurement per subject retained in the training set. The methods and settings are those of Table 1.

6.2 Results

Table 7 and Figure 5 report test MSE, and Table 8 presents paired comparisons.

Soybean.

Across all evaluated methods, MEXGB achieved the lowest mean squared error (MSE; 0.0383), followed by GPBoost (0.0395), GBM (0.0399), XGBoost (0.0406), and MEGB

Table 7 Applications: test MSE over 100 within-subject train/test splits, mean (standard error) and median. The lowest mean MSE for each dataset is shown in bold; – indicates that the method fitting was not available. Soybean: response on the log scale.

	BodyWeight		Soybean		Sprockett	
Method	Mean (SE)	Median	Mean (SE)	Median	Mean (SE)	Median
MEXGB	26.22 (0.70)	25.37	**0.0383 (0.0004)**	0.0387	0.176 (0.002)	0.172
MEGB	25.14 (0.64)	24.66	0.0425 (0.0005)	0.0424	0.194 (0.002)	0.192
GPBoost	**25.10 (0.57)**	24.95	0.0395 (0.0005)	0.0399	14.594 (10.994)	0.275
MERF	26.32 (0.95)	25.60	0.3325 (0.0061)	0.3330	0.182 (0.002)	0.180
REEMforest	25.19 (0.65)	25.48	0.3146 (0.0061)	0.3106	0.181 (0.002)	0.179
XGBoost	1641.97 (17.74)	1631.90	0.0406 (0.0005)	0.0409	0.172 (0.002)	0.170
GBM	1614.99 (18.09)	1587.32	0.0399 (0.0004)	0.0401	0.181 (0.002)	0.178
RF	3370.77 (26.37)	3360.91	1.0099 (0.0120)	1.0062	**0.166 (0.002)**	0.164
mboost	1189.89 (9.38)	1187.74	0.3302 (0.0029)	0.3317	1.217 (0.248)	0.444
glmmLasso	107.91 (11.91)	79.85	0.3351 (0.0030)	0.3342	–	–
LMM	77.25 (1.57)	77.22	0.3302 (0.0029)	0.3317	–	–

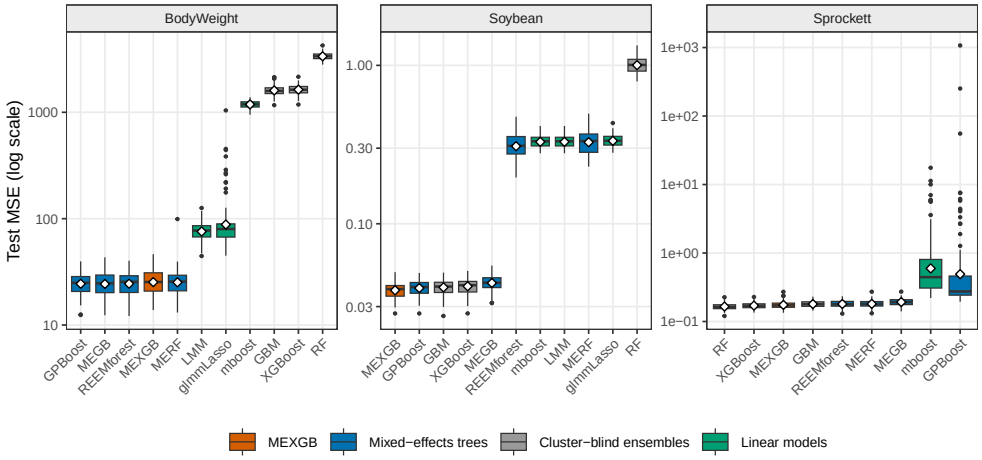


Fig. 5 Test MSE (logarithmic scale) over the 100 within-subject train/test splits of the three datasets. Within each panel the methods are ordered by mean MSE (diamonds), with the most accurate method on the left.

(0.0425). At the split level, MEXGB yielded lower error than GPBoost in 70% of splits and than MEGB in 89% of splits; the paired Wilcoxon test comparing MEXGB and GPBoost indicated a statistically significant difference ($p < 0.001$). The log-transformed leaf weight exhibits an approximately smooth temporal trajectory, with systematic heterogeneity across genotypes and years. In the proposed formulation, the population-level time trend is parameterized by $\boldsymbol{\beta}_z$ and a linear component, while the tree-based component captures nonlinear curvature; plot-level effects correspond to small deviations around this mean curve. By contrast, LMM (MSE 0.330) is unable to represent the observed curvature, and the mixed-effects forests (MSE 0.31–0.33) do not adequately recover this temporal pattern.

Table 8 Applications: paired comparison of MEXGB with the best competitor, MEGB and XGBoost on identical splits. Columns as in Table 3.

Data	Competitor	Diff.	Rel.	Better	p
BodyWeight	GPBoost	1.12	4	50	0.163
	MEGB	1.08	4	36	0.002
	XGBoost	-1615.75	-98	100	< 0.001
Soybean	GPBoost	-0.0013	-3	70	< 0.001
	MEGB	-0.0043	-10	89	< 0.001
	XGBoost	-0.0023	-6	84	< 0.001
Sprockett	RF	0.010	6	24	< 0.001
	MEGB	-0.018	-9	92	< 0.001
	XGBoost	0.004	3	34	< 0.001

BodyWeight.

All mixed-effects tree–based methods yielded comparably low prediction error, with MSE values in the range 25–27: GPBoost 25.10, MEGB 25.14, REEMforest 25.19, MEXGB 26.22, and MERF 26.32. In the split-level Wilcoxon comparison, the difference between MEXGB and GPBoost was not statistically significant ($p = 0.16$). By contrast, MEGB achieved a modestly lower mean error than MEXGB (approximately 4%), with strong evidence for a difference at the split level ($p = 0.002$).

In comparison, cluster-blind methods exhibited substantially larger MSEs (1190–3400), reflecting that between-rat heterogeneity dominates the measurement noise. This result underscores the advantage of explicitly modeling subject-specific variation when $\mathbf{D}$ is large. Given 16 rats and three predictors, two of which are constant within rat, the fixed-effects component contains limited information beyond the diet-specific means and the temporal trend. Consequently, the primary distinctions between methods arise from the way they estimate rat-specific intercepts and slopes for each of the 16 subjects.

Sprockett.

In terms of mean squared error (MSE), the RF (0.166) and XGBoost (0.172) exhibit the highest predictive accuracy, followed by MEXGB (0.176), GBM, REEMforest and MERF (0.181–0.182), and MEGB (0.194). Among the mixed-effects approaches, MEXGB yields the lowest error, outperforming MEGB in 92% of the data splits; however, it underperforms relative to the random forest in 76% of the splits. GPBoost occasionally produced extremely large errors on a small number of splits, thereby increasing its mean MSE (median 0.275). The MSE for MEXGB is comparable to that of its corresponding base learner, XGBoost (0.176 vs. 0.172), which suggests that incorporating random effects provides limited incremental predictive benefit in this setting. Table C3 in Appendix C further indicates that three of the four ASVs most frequently selected by MEXGB (a *Parabacteroides*, a *Prevotella*, and a *Ruminococcaceae* UCG-010 variant) are also selected by MEGB, XGBoost, and GBM in 68–86% of the splits, demonstrating substantial concordance in the dominant predictors across methods irrespective of whether random effects are included.

7 Discussion

MEXGB combines second-order tree boosting with a partially linear fixed effect and cross-fitted updates of the random effects. It achieved the lowest mean prediction error in all 13 simulation scenarios. Its variable-selection performance was strong in linear scenarios and broadly comparable with that of the other boosting methods in nonlinear scenarios. In the estimation study, MEXGB estimated the fixed effect more accurately than MEGB in every replicate, while the relative accuracy of variance-component estimation and random-effect prediction varied by scenario. The reduced-model comparisons highlight the contribution of the linear component for linear signals and suggest an important role for data-driven ensemble size in nonlinear settings.

Contributions of the model components.

The largest gains over tree-only mixed-effects methods occurred with linear signals: MEGB, MERF, and REEMforest had approximately four to fifteen times the prediction error of MEXGB. Theorems 5 and 6 illustrate why estimating a linear component directly can be more efficient than approximating it using step functions. This pattern is consistent with the broader finding in mixed-effects boosting that gains over tree-only alternatives concentrate in settings where the fixed-effect signal has a strong parametric component that trees approximate inefficiently, while in strongly nonlinear or strongly correlated settings the advantage narrows or reverses depending on how random effects are handled (Griesbach et al. 2021; Knieper et al. 2025; Olaniran et al. 2025). The partially linear structure also accommodates signals containing both linear and nonlinear components. The linear fit captures an initial part of the signal, and the trees model the remaining variation.

Cross-fitting and the comparison with other mixed-effects boosters.

Cross-fitting is intended to reduce the absorption of random-effect variation by a flexible fixed-effect learner. Proposition 1 establishes covariance collapse for an idealised interpolating learner, while the estimation study illustrates related finite-sample behaviour without complete interpolation. This mechanism distinguishes MEXGB from joint marginal-likelihood approaches such as GPBoost (Sigrist 2022, 2023), which estimate the tree ensemble and the random-effect covariance simultaneously rather than through cross-fitted EM/BLUP alternation. It also differs from the group-aware cross-validation strategy developed for LMM-XGBoost (Cho and Mueller 2026), which respects multilevel dependence in hyperparameter tuning and importance estimation but retains an in-sample estimation step for the fixed-effect component. The cross-fitting principle itself, however, is not specific to mixed models: in panel-data double/debiased machine learning, standard cross-fitting assumes independent observations and has been shown to require explicit adaptation when unobserved heterogeneity is present (Fuhr and Papies 2024). MEXGB can be read as one such adaptation for the longitudinal prediction task, where the target is not a causal coefficient but the conditional mean for a previously observed subject.

For squared-error loss, the Newton step differs from first-order least-squares boosting primarily through leaf-weight regularisation because $h_{ij} = 1$. Its broader motivation is the possibility of extending the framework to likelihoods with observation-specific Hessians. This is also the route through which MEXGB connects to recent generalized mixed-effects XGBoost proposals that accommodate continuous, binary, and count outcomes within a unified boosting framework (Asadi et al. 2026); the present implementation is deliberately restricted to continuous outcomes so that the variance-component and random-effect theory can be developed without an approximate E-step.

Interpretation of the comparisons.

The linear scenarios illustrate the benefit of an explicit linear component, although the LMM prediction benchmark also differs from MEXGB by omitting random slopes. The approximation bound in Theorem 5 is positive for a finite tree ensemble under its assumptions but decreases as the number of split points increases. It should therefore not be interpreted as a non-vanishing error floor for arbitrarily large ensembles. In the nonlinear scenarios, comparing full MEXGB with its trees-only version assesses the contribution of the linear component while retaining the other MEXGB procedures. It does not isolate every source of the difference from MEGB. The more general conclusion from the mixed-effects boosting literature is that the relative ranking of joint-estimation and alternating-estimation approaches depends on the strength of the correlation structure and on whether the marginal-likelihood optimisation is given adequate computational resources (Sigrist 2023; Knieper et al. 2025; Prevett et al. 2025).

Limitations.

MEXGB is more computationally demanding than a single booster because each outer iteration requires multiple boosting fits. With $p = 2000$, a fit takes approximately 9–15 minutes on one core. The fold models can be fitted in parallel. Finally, the theoretical results establish variance-component consistency conditional on fixed-effect consistency, an oracle inequality under specified fold-training assumptions, and lower bounds under particular model and design conditions. They do not establish a convergence rate for the complete adaptive MEXGB procedure. Extending existing boosting theory (Zhang and Yu 2005; Bühlmann and van de Geer 2011) to this setting remains an open problem. A further limitation shared with other mixed-effects tree methods is that variable importance is computed conditional on the fitted random-effect structure; recent work has shown that group-aware importance measures can reduce bias in multilevel designs (Cho and Mueller 2026), and adapting such measures to the cross-fitted MEXGB output is a natural next step.

When random effects provide little benefit.

The Sprockett application illustrates a setting in which the response depends almost entirely on the fixed component, so mixed-effects models offer very little contribution. MEXGB was close to XGBoost in predictive accuracy and was the most accurate mixed-effects method, but random forests had the lowest mean MSE overall. This

is consistent with the broader observation that cluster-blind ensembles can be competitive when between-subject heterogeneity is small relative to measurement-level variation, and that the incremental value of random effects is itself a quantity worth reporting rather than assuming (Capitaine et al. 2021; Olaniran et al. 2025).

Extensions.

The Newton boosting step can accommodate other likelihoods through their gradients and Hessians. Extending the full mixed-effects procedure to non-Gaussian outcomes would also require an appropriate, generally approximate, expectation step. Recent proposals that embed generalized linear mixed models within XGBoost provide a template for this direction (Asadi et al. 2026). Binary outcomes and clustered event times are the subject of separate work. Other extensions include structured residual covariance models, prediction intervals accounting for random-effect uncertainty, and dynamic prediction for new subjects as their first measurements become available.

8 Conclusion

We proposed MEXGB for high-dimensional longitudinal data with continuous outcomes. The method combines a relaxed lasso with Newton-boosted trees, uses cross-fitted fixed-effect predictions to update random effects, and selects model complexity using out-of-fold prediction error. Theoretical results motivate these components and describe their properties under specified assumptions. MEXGB achieved the lowest mean test error across 13 simulation scenarios with up to 2000 predictors. In the six main scenarios, it reduced test MSE by 13–83% relative to MEGB and by 42–95% relative to XGBoost. Across three applications, it was the most accurate method on one dataset, performed similarly to the leading mixed-effects methods on another, and was the most accurate mixed-effects method on the third.

These findings support MEXGB as a flexible option for predicting additional measurements from previously observed subjects. At the same time, the comparison with recent mixed-effects boosting work suggests that no single alternating or joint-estimation strategy dominates across all correlation regimes: MEXGB's advantage is clearest when the fixed-effect signal has an estimable linear component and when cross-fitting prevents the booster from absorbing subject-level variation, whereas joint marginal-likelihood approaches (Sigrist 2023; Prevett et al. 2025) and generalized mixed-effects XGBoost frameworks (Asadi et al. 2026; Cho and Mueller 2026) may be preferable when the random-effect variance is large or when non-Gaussian outcomes are of primary interest. The method is implemented in the R package `MEXGB`.

Declarations

- **Funding.** Not Applicable.
- **Competing interests.** The authors declare that they have no competing interests.
- **Data availability:** BodyWeight and Soybean are distributed with the R package `nlme`. The Sprockett data are available from the sources cited in Sprockett et al. (2020).

- **Ethics approval and consent to participate.** Not applicable. This study involved no human participants, human data, or animal subjects.
- **Consent for publication.** Not applicable.
- **Materials availability.** Not applicable.
- **Code availability.** The `MEXGB` R package (version 1.0.0) is freely available on GitHub at https://github.com/rid4stat/MEXGB. All simulation and data analysis codes required to reproduce the results presented in this paper are available from https://github.com/rid4stat/MEXGB-validation.
- **Author contributions.** ORO and SKA designed the simulation study. ORO and SKA conducted the simulation study, analysed the results, and drafted the manuscript. ORO, SKA, and SAA developed the `MEXGB` package. SAA, AB, and NMA contributed to the conceptualisation and study design, interpretation of results, and critical revision of the manuscript. ORO and SAA provided oversight of the project. All authors read and approved the final manuscript.

Appendix A Proofs

A.1 Auxiliary results

Lemma 1 (Newton step: leaf weights and split gain) *For a fixed tree structure with leaves $I_1, \dots, I_L$, write $G_\ell = \sum_{(i,j)\in I_\ell} g_{ij}$ and $H_\ell = \sum_{(i,j)\in I_\ell} h_{ij}$, and assume $H_\ell + \lambda > 0$. The objective (5) is minimised at $w_\ell^\star = -G_\ell/(H_\ell+\lambda)$, with minimum $-\frac{1}{2}\sum_\ell G_\ell^2/(H_\ell+\lambda)+\gamma L$, and splitting a node into two children lowers the minimum by (6). For the squared-error loss $L(y,\eta) = \frac{1}{2}(y-\eta)^2$, $g_{ij} = -u_{ij}$ with u_{ij} the current residual, $h_{ij} = 1$, and $w_\ell^\star = \{|I_\ell|/(|I_\ell|+\lambda)\}\,\bar{u}_\ell$, where $\bar{u}_\ell$ is the mean residual in leaf ℓ. With $\nu \in (0,1]$ the step lowers the penalty-free surrogate $\sum_{ij}\{g_{ij}\nu t(\mathbf{x}_{ij}) + \frac{1}{2}h_{ij}\nu^2 t(\mathbf{x}_{ij})^2\}$ by at least $\frac{1}{2}\nu(2-\nu)\sum_\ell G_\ell^2/(H_\ell+\lambda)$.*

Proof A tree with leaf weights $w_1, \dots, w_L$ takes the value w_ℓ at each measurement in I_ℓ, and every measurement belongs to exactly one leaf. Hence the objective (5) equals

$$\sum_{\ell=1}^{L}\Big\{w_\ell \sum_{I_\ell} g_{ij} + \tfrac{1}{2}w_\ell^2 \sum_{I_\ell} h_{ij}\Big\} + \gamma L + \frac{\lambda}{2}\sum_\ell w_\ell^2 = \sum_{\ell=1}^{L}\Big\{G_\ell w_\ell + \tfrac{1}{2}(H_\ell+\lambda)w_\ell^2\Big\} + \gamma L,$$

which separates into a sum of functions of the individual leaf weights w_ℓ. Each summand is a quadratic with positive leading coefficient $\frac{1}{2}(H_\ell+\lambda)$. It is therefore minimised where its derivative $G_\ell + (H_\ell+\lambda)w_\ell$ vanishes, at $w_\ell^\star = -G_\ell/(H_\ell+\lambda)$. Substituting, $G_\ell w_\ell^\star + \frac{1}{2}(H_\ell+\lambda)w_\ell^{\star 2} = -G_\ell^2/(H_\ell+\lambda) + \frac{1}{2}G_\ell^2/(H_\ell+\lambda) = -\frac{1}{2}G_\ell^2/(H_\ell+\lambda)$. An unsplit node I contributes $-\frac{1}{2}(G_\mathrm{L}+G_\mathrm{R})^2/(H_\mathrm{L}+H_\mathrm{R}+\lambda)+\gamma$ to the minimum; after splitting into I_L and I_R, it contributes $-\frac{1}{2}\{G_\mathrm{L}^2/(H_\mathrm{L}+\lambda) + G_\mathrm{R}^2/(H_\mathrm{R}+\lambda)\} + 2\gamma$. The difference is (6).

For the squared-error loss $\partial_\eta \frac{1}{2}(y-\eta)^2 = -(y-\eta)$ and $\partial_\eta^2 = 1$, so $g_{ij} = -u_{ij}$, $h_{ij} = 1$, $G_\ell = -|I_\ell|\bar{u}_\ell$, $H_\ell = |I_\ell|$ and $w_\ell^\star = |I_\ell|\bar{u}_\ell/(|I_\ell|+\lambda)$.

For the descent bound, the change of the surrogate on leaf ℓ is

$$\begin{aligned}\nu w_\ell^\star G_\ell + \tfrac{1}{2}\nu^2 w_\ell^{\star 2} H_\ell &= -\nu\frac{G_\ell^2}{H_\ell+\lambda} + \frac{\nu^2}{2}\frac{H_\ell G_\ell^2}{(H_\ell+\lambda)^2} \\ &\le -\nu\frac{G_\ell^2}{H_\ell+\lambda} + \frac{\nu^2}{2}\frac{G_\ell^2}{H_\ell+\lambda} = -\frac{\nu(2-\nu)}{2}\frac{G_\ell^2}{H_\ell+\lambda},\end{aligned}$$

because $H_\ell \le H_\ell + \lambda$. Summing over the leaves gives the bound, which is non-positive since $\nu(2-\nu) > 0$ for $\nu \in (0, 1]$. □

Lemma 2 (Expectation step) *Let* $\mathbf{r}_i = \mathbf{y}_i - \mathbf{f}_i - \mathbf{Z}_i\boldsymbol{\beta}_z$ *with* $(f, \boldsymbol{\beta}_z)$ *known and* $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \sigma^2\mathbf{I}$. *Then* $\mathbf{b}_i \mid \mathbf{y}_i \sim \mathcal{N}_q(\hat{\mathbf{b}}_i, \mathbf{C}_i)$ *with* $\hat{\mathbf{b}}_i$ *and* $\mathbf{C}_i$ *as in* (7), *and*

$$\mathbf{D}\mathbf{Z}_i^\top\mathbf{V}_i^{-1} = \big(\mathbf{D}^{-1} + \sigma^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\big)^{-1}\sigma^{-2}\mathbf{Z}_i^\top, \qquad \mathbf{D} - \mathbf{D}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D} = \big(\mathbf{D}^{-1} + \sigma^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\big)^{-1} \tag{A1}$$

when $\mathbf{D}$ *is invertible. Given* $(\hat{\mathbf{b}}_i, \mathbf{C}_i)$, *the maximisers of the EM surrogate* $Q(\mathbf{D}, \sigma^2) = \sum_i \mathbb{E}\{\log p(\mathbf{y}_i, \mathbf{b}_i) \mid \mathbf{y}_i\}$ *are given by* (8).

Proof The vector $(\mathbf{b}_i^\top, \mathbf{r}_i^\top)^\top$ is Gaussian with mean zero, $\operatorname{Var}\mathbf{b}_i = \mathbf{D}$, $\operatorname{Cov}(\mathbf{b}_i, \mathbf{r}_i) = \mathbf{D}\mathbf{Z}_i^\top$ and $\operatorname{Var}\mathbf{r}_i = \mathbf{V}_i$. The conditional distribution of a Gaussian vector gives $\mathbb{E}(\mathbf{b}_i \mid \mathbf{r}_i) = \mathbf{D}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}\mathbf{r}_i$ and $\operatorname{Var}(\mathbf{b}_i \mid \mathbf{r}_i) = \mathbf{D} - \mathbf{D}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}$, Conditioning on $\mathbf{y}_i$ is equivalent to conditioning on $\mathbf{r}_i$ when the fixed effect is known. To establish (A1), write $\mathbf{P} = \mathbf{D}^{-1} + \sigma^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i$. Then $\mathbf{P}\mathbf{D}\mathbf{Z}_i^\top = \mathbf{Z}_i^\top + \sigma^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top = \sigma^{-2}\mathbf{Z}_i^\top(\sigma^2\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top) = \sigma^{-2}\mathbf{Z}_i^\top\mathbf{V}_i$; multiplying on the left by $\mathbf{P}^{-1}$ and on the right by $\mathbf{V}_i^{-1}$ gives the first identity. By the first identity, $\mathbf{D}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D} = \mathbf{P}^{-1}\sigma^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\mathbf{D}$, so $\mathbf{D} - \mathbf{D}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D} = \mathbf{P}^{-1}(\mathbf{P}\mathbf{D} - \sigma^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\mathbf{D}) = \mathbf{P}^{-1}$.

For the EM step, $\log p(\mathbf{y}_i, \mathbf{b}_i) = \log p(\mathbf{y}_i \mid \mathbf{b}_i) + \log p(\mathbf{b}_i)$, and up to constants

$$Q = -\frac{n}{2}\log|\mathbf{D}| - \frac{1}{2}\operatorname{tr}\Big\{\mathbf{D}^{-1}\sum_i \mathbb{E}(\mathbf{b}_i\mathbf{b}_i^\top \mid \mathbf{y}_i)\Big\} - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i \mathbb{E}\big(\|\mathbf{r}_i - \mathbf{Z}_i\mathbf{b}_i\|^2 \mid \mathbf{y}_i\big).$$

The $\mathbf{D}$-part is $\frac{n}{2}\log|\mathbf{D}^{-1}| - \frac{1}{2}\operatorname{tr}(\mathbf{D}^{-1}\mathbf{S})$ with $\mathbf{S} = \sum_i \mathbb{E}(\mathbf{b}_i\mathbf{b}_i^\top \mid \mathbf{y}_i)$. This expression is concave in $\mathbf{D}^{-1}$, and its gradient $\frac{n}{2}\mathbf{D} - \frac{1}{2}\mathbf{S}$ vanishes at $\mathbf{D} = \mathbf{S}/n$. Since $\mathbb{E}(\mathbf{b}_i\mathbf{b}_i^\top \mid \mathbf{y}_i) = \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^\top + \mathbf{C}_i$, this is the first formula in (8), a sum of positive semidefinite matrices. The part involving σ^2 is maximised at $\sigma^2 = N^{-1}\sum_i \mathbb{E}(\|\mathbf{r}_i - \mathbf{Z}_i\mathbf{b}_i\|^2 \mid \mathbf{y}_i)$, and $\mathbb{E}(\|\mathbf{r}_i - \mathbf{Z}_i\mathbf{b}_i\|^2 \mid \mathbf{y}_i) = \|\mathbf{r}_i - \mathbf{Z}_i\hat{\mathbf{b}}_i\|^2 + \operatorname{tr}(\mathbf{Z}_i\mathbf{C}_i\mathbf{Z}_i^\top)$ because $\mathbf{b}_i - \hat{\mathbf{b}}_i$ has conditional mean zero and covariance $\mathbf{C}_i$. □

Lemma 3 (Centring of the random effects) *For any* $\mathbf{m} \in \mathbb{R}^q$, $(\boldsymbol{\beta}_z, \{\mathbf{b}_i\})$ *and* $(\boldsymbol{\beta}_z + \mathbf{m}, \{\mathbf{b}_i - \mathbf{m}\})$ *give the same linear predictors. For predicted random effects with mean* $\bar{\mathbf{b}}$,

$$\frac{1}{n}\sum_i \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^\top = \frac{1}{n}\sum_i (\hat{\mathbf{b}}_i - \bar{\mathbf{b}})(\hat{\mathbf{b}}_i - \bar{\mathbf{b}})^\top + \bar{\mathbf{b}}\bar{\mathbf{b}}^\top.$$

Hence, if a population trend $\mathbf{z}^\top\mathbf{m}$ *is represented by neither* f *nor* $\boldsymbol{\beta}_z$, *it can enter* $\bar{\mathbf{b}} \approx \mathbf{m}$ *and adds* $\mathbf{m}\mathbf{m}^\top$ *to the covariance update* (8); *the centring* $\boldsymbol{\beta}_z \leftarrow \boldsymbol{\beta}_z + \bar{\mathbf{b}}$, $\hat{\mathbf{b}}_i \leftarrow \hat{\mathbf{b}}_i - \bar{\mathbf{b}}$ *removes this term without changing any linear predictor.*

Proof $\mathbf{z}^\top(\boldsymbol{\beta}_z + \mathbf{m}) + \mathbf{z}^\top(\mathbf{b}_i - \mathbf{m}) = \mathbf{z}^\top\boldsymbol{\beta}_z + \mathbf{z}^\top\mathbf{b}_i$. To obtain the identity, expand $\hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^\top = \{(\hat{\mathbf{b}}_i - \bar{\mathbf{b}}) + \bar{\mathbf{b}}\}\{(\hat{\mathbf{b}}_i - \bar{\mathbf{b}}) + \bar{\mathbf{b}}\}^\top$ and average; the cross terms vanish because $\sum_i(\hat{\mathbf{b}}_i - \bar{\mathbf{b}}) = \mathbf{0}$. □

Lemma 4 (Restriction of the linear component) *Let* $\ell_{\min} \le \ell_{\max}$ *be the smallest and largest fitted values of the linear component on the training data and* $\Pi(u) = \min\{\max(u, \ell_{\min}), \ell_{\max}\}$. *For every* $u \in \mathbb{R}$ *and every* $c \in [\ell_{\min}, \ell_{\max}]$, $|\Pi(u) - c| \le |u - c|$.

Proof If $u \in [\ell_{\min}, \ell_{\max}]$, $\Pi(u) = u$. If $u > \ell_{\max}$, $\Pi(u) = \ell_{\max}$ and $|\ell_{\max} - c| = \ell_{\max} - c < u - c = |u - c|$ because $c \leq \ell_{\max} < u$. The case $u < \ell_{\min}$ is symmetric. □

Lemma 5 (Cost of growing the fold models in chunks) *Suppose that continuing a boosted model of m trees requires a copy of the model at a cost proportional to m. Growing M trees in chunks of fixed size c costs $O(M^2/c)$ for the copies; chunks whose size equals the current model size (doubling), capped at $c_{\max}$, cost $O(M)$ when $M \leq c_{\max}$ and $O(M + M^2/c_{\max})$ otherwise. Evaluating only the trees added in each chunk on the held-out measurements costs one evaluation per tree and held-out measurement in total.*

Proof With fixed-size chunks, there are M/c continuations and the k-th continuation copies kc trees, for a total of $c\sum_{k \leq M/c} k = O(M^2/c)$. With doubling chunks, the numbers of copied trees form a geometric series whose sum is at most $2M$; once the cap is reached the remaining $(M - c_{\max})/c_{\max}$ continuations copy at most M trees each. The out-of-fold prediction margin is updated cumulatively, so each tree is evaluated once on each held-out measurement. □

Proposition 10 (Monotonicity of the in-sample algorithm) *Consider the in-sample version of the algorithm, in which the expectation step uses $\hat{f}^{\text{full}}$, with a fixed structure of f. If each update of f does not decrease $Q(\cdot \mid \boldsymbol{\theta}^{(t-1)})$ and the variance components are updated by (8), then $\ell(\boldsymbol{\theta}^{(t)}) \geq \ell(\boldsymbol{\theta}^{(t-1)})$, and the sequence $\ell(\boldsymbol{\theta}^{(t)})$ converges if ℓ is bounded above.*

Proof For any $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, $\ell(\boldsymbol{\theta}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') - H(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$ with $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = \sum_i \mathbb{E}_{\mathbf{b}_i \mid \mathbf{y}_i, \boldsymbol{\theta}'} \log p(\mathbf{b}_i \mid \mathbf{y}_i, \boldsymbol{\theta})$. By Gibbs' inequality $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}') \leq H(\boldsymbol{\theta}' \mid \boldsymbol{\theta}')$ (Dempster et al. 1977). Hence

$$\begin{aligned}\ell(\boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta}^{(t-1)}) = {} & \left\{Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)}) - Q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^{(t-1)})\right\} \\ & + \left\{H(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^{(t-1)}) - H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t-1)})\right\} \geq 0.\end{aligned}$$

The first bracketed term is non-negative because the update of f does not decrease Q, the update (8) maximises Q in $(\mathbf{D}, \sigma^2)$ (Lemma 2) and the centring leaves ℓ and Q unchanged (Lemma 3); the second is non-negative by Gibbs' inequality. A non-decreasing sequence that is bounded above converges. □

A.2 Proofs of the results of Section 4.1

Proof of Proposition 1 At iteration t the in-sample residual of every measurement of subject i is $y_{ij} - \hat{f}^{(t)}(\mathbf{x}_{ij}) = \hat{b}_i^{(t-1)}$, so the residual vector is $\hat{b}_i^{(t-1)}\mathbf{1}_m$. With $\mathbf{V} = d_t \mathbf{1}\mathbf{1}^\top + \sigma^2 \mathbf{I}$, the Sherman–Morrison formula gives $\mathbf{V}^{-1}\mathbf{1} = \mathbf{1}/(\sigma^2 + m d_t)$, and (7) yields

$$\hat{b}_i^{(t)} = d_t \mathbf{1}^\top \mathbf{V}^{-1} \mathbf{1}\, \hat{b}_i^{(t-1)} = \frac{m d_t}{m d_t + \sigma^2} \hat{b}_i^{(t-1)} = k_t \hat{b}_i^{(t-1)}, \qquad C_t = d_t - \frac{m d_t^2}{m d_t + \sigma^2} = d_t(1 - k_t).$$

Write $\beta_t = n^{-1} \sum_i (\hat{b}_i^{(t-1)})^2$. Then $n^{-1} \sum_i (\hat{b}_i^{(t)})^2 = k_t^2 \beta_t = \beta_{t+1}$, and (8) gives

$$d_{t+1} = k_t^2 \beta_t + d_t(1 - k_t). \tag{A2}$$

Since $k_t < 1$, $\beta_{t+1} \leq \beta_t$. For $t \geq 2$, (A2) at $t-1$ gives $d_t \geq k_{t-1}^2 \beta_{t-1} = \beta_t$, hence $d_{t+1} - d_t = k_t(k_t \beta_t - d_t) \leq k_t(\beta_t - d_t) \leq 0$. Both sequences are non-negative and non-increasing from the stated iteration onwards, so they converge: $d_t \to d_\infty \geq 0$ and $\beta_t \to \beta_\infty \geq 0$. Suppose

$d_\infty > 0$. Then $k_t \to k_\infty = md_\infty/(md_\infty + \sigma^2) \in (0,1)$; passing to the limit in $\beta_{t+1} = k_t^2 \beta_t$ gives $\beta_\infty = k_\infty^2 \beta_\infty$, so $\beta_\infty = 0$. Taking the limit in (A2) then gives $d_\infty = d_\infty(1 - k_\infty)$, so $k_\infty d_\infty = 0$, a contradiction. Therefore $d_\infty = 0$. Finally, $\beta_t \to 0$ implies $n^{-1}\sum_i (\hat{b}_i^{(t)})^2 \to 0$. With n fixed, this implies $\hat{b}_i^{(t)} \to 0$ for every subject. □

Proof of Proposition 2 With a random intercept, $y_{kl} = f(\mathbf{x}_{kl}) + b_k + \varepsilon_{kl}$, and

$$y_{ij} - \sum_{k,l} S_{ij,kl} y_{kl} = b_i \Big(1 - \sum_l S_{ij,il}\Big) - \sum_{k \neq i} b_k \sum_l S_{ij,kl} + \Big\{ f(\mathbf{x}_{ij}) + \varepsilon_{ij} - \sum_{k,l} S_{ij,kl} \big(f(\mathbf{x}_{kl}) + \varepsilon_{kl}\big) \Big\}.$$

The random effects b_k, $k \neq i$, are independent of b_i, and the final bracketed term contains no random effects, so b_i enters with coefficient $1 - \pi_{ij}$. For the cross-fitted smoother, the row corresponding to $(i,j) \in \mathcal{F}_v$ assigns zero weight to $\mathcal{F}_v$ and total weight one to the $|\mathcal{F}_v^c|$ measurements of $\mathcal{F}_v^c$. If the measurements are exchangeable and the weights do not depend on the subject labels, each of these measurements receives expected weight $1/|\mathcal{F}_v^c|$. With folds balanced within subjects, $\mathcal{F}_v^c$ contains $n_i(1 - 1/V)$ measurements of subject i and $|\mathcal{F}_v^c| = N(1 - 1/V)$, so $\mathbb{E}\pi_{ij}^{\mathrm{cf}} = n_i(1 - 1/V)/\{N(1 - 1/V)\} = n_i/N$. □

Proof of Proposition 3 The decomposition follows from (1) and the definition of $\mathbf{e}_i$. For $(i,j) \in \mathcal{F}_v$, $\hat{f}^{\mathrm{cf}}(\mathbf{x}_{ij}) = \hat{f}^{(-v)}(\mathbf{x}_{ij})$ is a function of $\mathbf{x}_{ij}$ and of $\mathcal{D}^{(-v)}$. The error ε_{ij} is independent of $\mathbf{x}_{ij}$, of the random effects and of all other errors, hence of $\mathcal{D}^{(-v)}$, and therefore of e_{ij} conditionally on $\mathcal{D}^{(-v)}$. Under the stated conditions, $\boldsymbol{\varepsilon}_i + \mathbf{e}_i$ has independent components with mean zero and variance $\sigma^2 + \tau^2$, independent of $\mathbf{b}_i$, so the residuals follow a linear mixed model with covariance $\mathbf{D}$ and error variance $\sigma^2 + \tau^2$. By Lemma 2 the updates (7)–(8) are the EM algorithm for this model, whose fixed points are its stationary points (the maximum likelihood estimate when it is interior and unique). With a Gaussian working model the estimate of $\mathbf{D}$ is unaffected by the extra error variance, which is absorbed by the estimate of the residual variance. □

Proof of Proposition 4 With $\mathbf{V}_i = d\mathbf{1}\mathbf{1}^\top + s^2\mathbf{I}$ and residual vector $\mathbf{r}_i$, the Sherman–Morrison formula gives $\hat{b}_i = d\mathbf{1}^\top \mathbf{r}_i/(s^2 + md) = d\bar{r}_i/(d+a)$ and $C_i = d - md^2/(s^2 + md) = da/(d+a)$, so the update (8) for the intercept variance is $g(d) = n^{-1}\sum_i (\hat{b}_i^2 + C_i) = d^2(d+a)^{-2}\,\overline{\bar{r}^2} + da/(d+a)$. The marginal likelihood depends on d only through the subject means: $\bar{r}_i \sim \mathcal{N}(0, d+a)$ independently, and the distribution of the within-subject deviations $r_{ij} - \bar{r}_i$ does not depend on d. Maximising $-\frac{1}{2}\sum_i \{\log(d+a) + \bar{r}_i^2/(d+a)\}$ gives $d^\star + a = \overline{\bar{r}^2}$. Substituting, $g(d^\star) = \{d^{\star 2} + d^\star a\}/(d^\star + a) = d^\star$. Differentiating,

$$g'(d) = \frac{2da}{(d+a)^3}\,\overline{\bar{r}^2} + \frac{a^2}{(d+a)^2}, \qquad g'(d^\star) = \frac{2d^\star a + a^2}{(d^\star + a)^2} = 1 - \frac{d^{\star 2}}{(d^\star + a)^2} = 1 - \rho^2.$$

For a map with $0 < g'(d^\star) < 1$, the error near $d^\star$ is approximately multiplied by $g'(d^\star)$ at every step. □

A.3 Proofs of the out-of-fold selection results

Proof of Theorem 1 Condition on $\mathcal{D}^{(-v)}$. The fold models and the held-out offsets are then fixed functions. Data from different subjects are independent, and the conditional law of the held-out measurements of subject i given $\mathcal{D}^{(-v)}$ depends only on the measurements of subject i in $\mathcal{D}^{(-v)}$; hence the held-out blocks of different subjects are conditionally independent. For a candidate $\mathfrak{f}$ let $S_i(\mathfrak{f})$ be the sum of the held-out squared errors of subject i; it

lies in $[0, Bm_{\max}]$, and $\widehat{\mathcal{L}}_v(\mathfrak{f}) = \sum_i S_i(\mathfrak{f})$. Hoeffding's inequality (Hoeffding 1963) for the n conditionally independent bounded variables gives, for each $\mathfrak{f}$ and $s > 0$,

$$\mathbb{P}\Big(\big|\widehat{\mathcal{L}}_v(\mathfrak{f}) - |\mathcal{F}_v| R_v(\mathfrak{f})\big| > s \;\Big|\; \mathcal{D}^{(-v)}\Big) \le 2\exp\Big(-\frac{2s^2}{nB^2 m_{\max}^2}\Big).$$

Choose $s = Bm_{\max}\{n\log(2|\mathcal{G}|/\delta)/2\}^{1/2}$, so that the right-hand side is $\delta/|\mathcal{G}|$, and apply the union bound over the $|\mathcal{G}|$ candidates: with conditional probability at least $1-\delta$ all deviations are at most s. On this event, with $\mathfrak{f}^\star$ a minimiser of R_v,

$$|\mathcal{F}_v| R_v(\hat{\mathfrak{f}}) \le \widehat{\mathcal{L}}_v(\hat{\mathfrak{f}}) + s \le \widehat{\mathcal{L}}_v(\mathfrak{f}^\star) + s \le |\mathcal{F}_v| R_v(\mathfrak{f}^\star) + 2s,$$

where the middle inequality holds because $\hat{\mathfrak{f}}$ minimises $\widehat{\mathcal{L}}_v$. Dividing by $|\mathcal{F}_v|$ establishes the bound. Integrating over $\mathcal{D}^{(-v)}$ gives the unconditional probability statement. □

Proof of Proposition 5 The predictor ranking of fold v is computed from the fold model fitted on $\mathcal{F}_v^c$ with the training offsets, and the screened fold model is refitted on $\mathcal{F}_v^c$ with the same offsets; both are functions of $\mathcal{D}^{(-v)}$ and of the offsets. Conditional on these quantities, each candidate is a fixed function, and the out-of-fold loss is a sum of evaluations of a fixed function at the held-out measurements, whose conditional expectation is $|\mathcal{F}_v|$ times the conditional risk. The conditions of Theorem 1 are therefore met for the finite set of screened candidates including the candidate using all predictors. If instead the ranking were computed on all measurements, it would depend on the held-out responses, and the loss of the screened candidates would be biased downwards. □

Lemma 6 (Gaussian maxima) *Let $Z_1, \dots, Z_r$ be possibly dependent standard normal random variables. Then $\mathbb{E}\max_k |Z_k| \le \{2\log(2r)\}^{1/2}$. If, in addition, $Z_k = \mathbf{a}_k^\top \mathbf{g}$ with $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\|\mathbf{a}_k\| = 1$, then $\mathrm{Var}(\max_k |Z_k|) \le 1$ and $\mathbb{E}\max_k Z_k^2 \le 2\log(2r) + 1$.*

Proof The $2r$ variables $\pm Z_k$ are standard normal and $\max_k |Z_k| = \max_{k,\pm}(\pm Z_k)$. For $s > 0$, Jensen's inequality gives $\exp\{s\mathbb{E}\max_k |Z_k|\} \le \mathbb{E}\exp\{s\max_{k,\pm}(\pm Z_k)\} \le \sum_{k,\pm}\mathbb{E}e^{\pm sZ_k} = 2re^{s^2/2}$, so $\mathbb{E}\max_k |Z_k| \le s^{-1}\log(2r) + s/2$, and $s = \{2\log(2r)\}^{1/2}$ gives the first claim. The function $F(\mathbf{g}) = \max_k |\mathbf{a}_k^\top \mathbf{g}|$ satisfies $|F(\mathbf{g}) - F(\mathbf{g}')| \le \max_k |\mathbf{a}_k^\top(\mathbf{g} - \mathbf{g}')| \le \|\mathbf{g} - \mathbf{g}'\|$, so it is 1-Lipschitz, and the Gaussian Poincaré inequality (Boucheron et al. 2013, Chapter 3) gives $\mathrm{Var}\, F(\mathbf{g}) \le 1$. Finally $\mathbb{E}\max_k Z_k^2 = \mathbb{E}F^2 = (\mathbb{E}F)^2 + \mathrm{Var}\, F \le 2\log(2r) + 1$. □

Proof of Proposition 6 With $h \equiv 1$, the Hessian sums equal the leaf sizes. A median split gives two children of size $m/2$, with $G_{\mathrm{L}} = -\sum_{\mathrm{L}} u_{ij}$ and $G_{\mathrm{R}} = -\sum_{\mathrm{R}} u_{ij}$, so with $\lambda = \gamma = 0$,

$$\mathrm{Gain} = \frac{1}{2}\Big\{\frac{2G_{\mathrm{L}}^2}{m} + \frac{2G_{\mathrm{R}}^2}{m} - \frac{(G_{\mathrm{L}} + G_{\mathrm{R}})^2}{m}\Big\} = \frac{(G_{\mathrm{L}} - G_{\mathrm{R}})^2}{2m}.$$

Because the predictor is independent of the residuals, the split is independent of $\mathbf{u}$, and $(G_{\mathrm{L}} - G_{\mathrm{R}})/(\sigma\sqrt{m}) = \mathbf{a}_k^\top \mathbf{g}$ with $\mathbf{g} = \mathbf{u}/\sigma \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ and $\mathbf{a}_k$ a vector with entries $\pm m^{-1/2}$, so $\|\mathbf{a}_k\| = 1$. Hence $\mathrm{Gain}_k = \frac{1}{2}\sigma^2 Z_k^2$ with $Z_k = \mathbf{a}_k^\top \mathbf{g} \sim \mathcal{N}(0,1)$, and Lemma 6 gives the bound. □

Proof of Proposition 7 We treat the fold models as fixed functions; their dependence on any single subject is of order $1/n$. Let $(i,j) \in \mathcal{F}_v$ and let $\mathcal{M}$ be the measurements of subject i outside $\mathcal{F}_v$. The predictor of y_{ij} is $\hat{f}^{(-v)}(\mathbf{x}_{ij}) + \mathbf{z}_{ij}^\top\{\boldsymbol{\beta}_z + \hat{\mathbf{b}}_i^{(-v)}(c\mathbf{D})\}$, a fixed function ψ of $(\mathbf{x}_{ij}, \mathbf{z}_{ij})$ and of the data in $\mathcal{M}$. For a new measurement $(\mathbf{x}^*, \mathbf{z}^*, y^*)$ of subject i the same predictor is $\psi(\mathbf{x}^*, \mathbf{z}^*; \mathcal{M})$. By exchangeability of the measurements of subject i given $\mathbf{b}_i$, the triples $\big((\mathbf{x}_{ij}, \mathbf{z}_{ij}, y_{ij}), \mathcal{M}\big)$ and $\big((\mathbf{x}^*, \mathbf{z}^*, y^*), \mathcal{M}\big)$ have the same joint law. Identical distributions imply identical expectations of the squared error $\{y - \psi(\mathbf{x}, \mathbf{z}; \mathcal{M})\}^2$. Summing over the held-out measurements and dividing by N establishes the stated unbiasedness of $\mathcal{E}(c)/N$. □

Proof of Proposition 8 With the true parameters, the held-out residual is $r_{ij} = y_{ij} - f(\mathbf{x}_{ij}) - \mathbf{z}_{ij}^\top\boldsymbol{\beta}_z = \mathbf{z}_{ij}^\top\mathbf{b}_i + \varepsilon_{ij}$. Since ε_{ij} is independent of $\mathbf{b}_i$ and of the residuals $\mathbf{r}_{i,-v}$ of the subject's other folds, $\mathbb{E}(r_{ij} \mid \mathbf{r}_{i,-v}) = \mathbf{z}_{ij}^\top\mathbb{E}(\mathbf{b}_i \mid \mathbf{r}_{i,-v}) = \mathbf{z}_{ij}^\top\hat{\mathbf{b}}_i^{(-v)}(\mathbf{D})$ by Lemma 2 applied to the measurements outside $\mathcal{F}_v$. The conditional mean minimises the expected squared error among all functions of $\mathbf{r}_{i,-v}$, and $\mathbf{z}_{ij}^\top\hat{\mathbf{b}}_i^{(-v)}(c\mathbf{D})$ is such a function for every c. Hence the expected squared error is minimised at $c = 1$ for every held-out measurement, and so is the expectation of the sum $\mathcal{E}(c)$. □

A.4 Proofs of the consistency results

Proof of Theorem 2 With the true offsets, $y - o - g(\mathbf{x}) = \{f(\mathbf{x}) - g(\mathbf{x})\} + \varepsilon$ with $\mathbb{E}(\varepsilon \mid \mathbf{x}, \mathbf{z}, \mathbf{b}) = 0$ and $\operatorname{Var}\varepsilon = \sigma^2$. Conditioning on $(\mathbf{x}, \mathbf{z}, \mathbf{b})$, the cross term has expectation zero, so $R(g) = \sigma^2 + \mathbb{E}\{f(\mathbf{x}) - g(\mathbf{x})\}^2$. Hence $R(g) - R(f) = \mathbb{E}\{g(\mathbf{x}) - f(\mathbf{x})\}^2 \geq 0$, with equality if and only if $g = f$ almost surely. □

Proof of Theorem 3 Write $\mathbf{r}_i = \mathbf{y}_i - \mathbf{f}_i - \mathbf{Z}_i\boldsymbol{\beta}_z = \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$ and $\mathbf{e}_i = \mathbf{f}_i - \hat{\mathbf{f}}_i$, so that $\mathbf{y}_i - \hat{\mathbf{f}}_i - \mathbf{Z}_i\boldsymbol{\beta}_z = \mathbf{r}_i + \mathbf{e}_i$. For every $\boldsymbol{\theta} \in \Theta$, the largest eigenvalue of $\mathbf{V}_i^{-1}$ is at most $\sigma_{\min}^{-2}$, hence

$$\big|(\mathbf{r}_i + \mathbf{e}_i)^\top\mathbf{V}_i^{-1}(\mathbf{r}_i + \mathbf{e}_i) - \mathbf{r}_i^\top\mathbf{V}_i^{-1}\mathbf{r}_i\big| = \big|2\mathbf{e}_i^\top\mathbf{V}_i^{-1}\mathbf{r}_i + \mathbf{e}_i^\top\mathbf{V}_i^{-1}\mathbf{e}_i\big| \leq \sigma_{\min}^{-2}\big(2\|\mathbf{r}_i\|\|\mathbf{e}_i\| + \|\mathbf{e}_i\|^2\big).$$

Write $\ell_n(\boldsymbol{\theta}; \tilde{f}) = n^{-1}\sum_i \ell_i(\boldsymbol{\theta}; \tilde{f})$. Averaging and applying the Cauchy–Schwarz inequality,

$$\sup_{\boldsymbol{\theta}\in\Theta}\big|\ell_n(\boldsymbol{\theta}; \hat{f}) - \ell_n(\boldsymbol{\theta}; f)\big| \leq \sigma_{\min}^{-2}\Big\{\Big(\frac{1}{n}\sum_i \|\mathbf{r}_i\|^2\Big)^{1/2}\Big(\frac{1}{n}\sum_i \|\mathbf{e}_i\|^2\Big)^{1/2} + \frac{1}{2n}\sum_i \|\mathbf{e}_i\|^2\Big\}.$$

By the law of large numbers $n^{-1}\sum_i \|\mathbf{r}_i\|^2 = O_p(1)$, since $\mathbb{E}\|\mathbf{r}_i\|^2 \leq m_{\max}\{\lambda_{\max}(\mathbf{Z}_i\mathbf{D}_0\mathbf{Z}_i^\top) + \sigma_0^2\}$ is bounded, and $n^{-1}\sum_i \|\mathbf{e}_i\|^2 = (N/n)\Delta_N \leq m_{\max}\Delta_N \to 0$ in probability. The perturbation therefore converges uniformly to zero in probability. The functions $\boldsymbol{\theta} \mapsto \ell_i(\boldsymbol{\theta}; f)$ are continuous on the compact set Θ and dominated by the integrable envelope $\frac{1}{2}\sup_\Theta|\log|\mathbf{V}_i|| + \frac{1}{2}\sigma_{\min}^{-2}\|\mathbf{r}_i\|^2$, so the uniform law of large numbers (van der Vaart 1998, Example 19.8) gives $\sup_\Theta|\ell_n(\boldsymbol{\theta}; f) - \mathbb{E}\ell_i(\boldsymbol{\theta}; f)| \to 0$. Together, $\sup_\Theta|\ell_n(\boldsymbol{\theta}; \hat{f}) - \mathbb{E}\ell_i(\boldsymbol{\theta}; f)| \to 0$. The limiting criterion is continuous on a compact set and has a unique, well-separated maximiser, and Theorem 5.7 of van der Vaart (1998) gives $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ in probability. □

A.5 Proofs of the comparison results

Proof of Theorem 4 Write $y^* - \hat{g} = A + B$ with $A = f(\mathbf{x}^*) + \mathbf{z}^{*\top}\boldsymbol{\beta}_z - \hat{g}$ and $B = \mathbf{z}^{*\top}\mathbf{b}_i + \varepsilon^*$. A is a function of $(\mathbf{x}^*, \mathbf{z}^*, \hat{g})$, and by hypothesis $\mathbb{E}(B \mid \mathbf{x}^*, \mathbf{z}^*, \hat{g}) = 0$, so $\mathbb{E}(AB) = 0$ and $\mathbb{E}(y^* - \hat{g})^2 = \mathbb{E}A^2 + \mathbb{E}B^2 \geq \mathbb{E}B^2$. By the independence of $\mathbf{b}_i$, ε^*, and $\mathbf{z}^*$, $\mathbb{E}B^2 = \mathbb{E}(\mathbf{z}^{*\top}\mathbf{D}\mathbf{z}^*) +$

σ^2. For the mixed-model predictor the error is $\mathbf{z}^{*\top}(\mathbf{b}_i - \hat{\mathbf{b}}_i) + \varepsilon^*$; by Lemma 2, $\mathbf{b}_i - \hat{\mathbf{b}}_i$ has conditional mean zero and covariance $\mathbf{C}_i$ given the training measurements of subject i, and ε^* is independent of these, so the mean squared error is $\sigma^2 + \mathbb{E}(\mathbf{z}^{*\top}\mathbf{C}_i\mathbf{z}^*)$. By (A1), $\mathbf{C}_i^{-1} = \mathbf{D}^{-1} + \sigma^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i \succeq \mathbf{D}^{-1}$, hence $\mathbf{C}_i \preceq \mathbf{D}$, and $\lambda_{\max}(\mathbf{C}_i) \le \sigma^2/\lambda_{\min}(\mathbf{Z}_i^\top\mathbf{Z}_i) \to 0$. □

Proof of Proposition 9 Write $\mathbf{r}_i = \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$, so that $\mathbf{y}_i - \hat{\mathbf{f}}_i - \mathbf{Z}_i\boldsymbol{\beta}_z = \mathbf{r}_i + \mathbf{e}_i$ and $\hat{\mathbf{b}}_i = \tilde{\mathbf{b}}_i + \mathbf{D}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}\mathbf{e}_i$ with $\tilde{\mathbf{b}}_i = \mathbf{D}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}\mathbf{r}_i$, the predictor based on the true residuals. The prediction error is

$$\begin{aligned} y^* - \hat{f}(\mathbf{x}^*) - \mathbf{z}^{*\top}(\boldsymbol{\beta}_z + \hat{\mathbf{b}}_i) &= e^* + \mathbf{z}^{*\top}\mathbf{b}_i + \varepsilon^* - \mathbf{z}^{*\top}\tilde{\mathbf{b}}_i - \mathbf{k}^\top\mathbf{e}_i \\ &= \underbrace{\{\mathbf{z}^{*\top}(\mathbf{b}_i - \tilde{\mathbf{b}}_i) + \varepsilon^*\}}_{U} + \underbrace{\{e^* - \mathbf{k}^\top\mathbf{e}_i\}}_{W}. \end{aligned}$$

Let $\mathcal{P}$ denote the predictors and random-effect designs of subject i for both training and new measurements, together with $\hat{f}$. W is a function of $\mathcal{P}$. Given $\mathcal{P}$, U depends only on $(\mathbf{b}_i, \boldsymbol{\varepsilon}_i, \varepsilon^*)$, which are independent of $\mathcal{P}$; by Lemma 2, $\mathbf{b}_i - \tilde{\mathbf{b}}_i$ has conditional mean zero and covariance $\mathbf{C}_i$, and ε^* is independent of it with variance σ^2. Hence $\mathbb{E}(U \mid \mathcal{P}) = 0$, $\mathbb{E}(UW) = 0$ and $\mathbb{E}(U^2 \mid \mathcal{P}) = \sigma^2 + \mathbf{z}^{*\top}\mathbf{C}_i\mathbf{z}^*$, which gives (17). □

Proof of Corollary 1 With $\hat{f} = a^\circ + \mathbf{x}^\top\boldsymbol{\beta}^\circ$, $e(\mathbf{x}) = f(\mathbf{x}) - a^\circ - \mathbf{x}^\top\boldsymbol{\beta}^\circ$. Because the projection includes an intercept, $\mathbb{E}e(\mathbf{x}) = 0$, and $\mathbb{E}e(\mathbf{x})^2 = \Delta_{\text{lin}}(f)$. The variable e^* is a function of $\mathbf{x}^*$, while $\mathbf{k}^\top\mathbf{e}_i$ is a function of the training predictors of subject i and of $\mathbf{z}^*$; they are independent under the stated assumptions. Hence $\mathbb{E}(e^* - \mathbf{k}^\top\mathbf{e}_i)^2 = \mathrm{Var}(e^*) + \mathbb{E}(\mathbf{k}^\top\mathbf{e}_i)^2 - 2\mathbb{E}(e^*)\mathbb{E}(\mathbf{k}^\top\mathbf{e}_i) = \Delta_{\text{lin}}(f) + \mathbb{E}(\mathbf{k}^\top\mathbf{e}_i)^2 \ge \Delta_{\text{lin}}(f)$, and Proposition 9 gives the claim. If f is not affine almost surely, the projection error $\Delta_{\text{lin}}(f)$ is positive. □

Proof of Theorem 5 Reduction to $\mathbf{x}_A$. Let $\bar{g}(\mathbf{x}_A) = \mathbb{E}\{g(\mathbf{x}) \mid \mathbf{x}_A\}$. Since f depends on $\mathbf{x}_A$ only, $\mathbb{E}(f-g)^2 = \mathbb{E}(f-\bar{g})^2 + \mathbb{E}(g-\bar{g})^2 \ge \mathbb{E}(f-\bar{g})^2$. Each tree is a sum of terms $w\mathbb{1}\{\mathbf{x}_A \in R_A\}\mathbb{1}\{\mathbf{x}_B \in R_B\}$, where R_A is a product of intervals whose endpoints are split points of the ensemble on $x_1, \dots, x_s$. By independence of $\mathbf{x}_A$ and $\mathbf{x}_B$, $\mathbb{E}\{\mathbb{1}(\mathbf{x}_A \in R_A)\mathbb{1}(\mathbf{x}_B \in R_B) \mid \mathbf{x}_A\} = \mathbb{1}(\mathbf{x}_A \in R_A)\,\mathbb{P}(\mathbf{x}_B \in R_B)$. Hence $\bar{g}$ is constant on the cells of the grid formed by the K_k split points on each coordinate x_k, $k \le s$.

Cell variance. The split points on x_k divide $[0,1]$ into $L_k = K_k + 1$ intervals of lengths $h_{k1}, \dots, h_{kL_k}$ summing to one, and the grid cells are products of these intervals. On a cell Q, $\bar{g}$ is constant, so $\mathbb{E}\{(f-\bar{g})^2 \mid \mathbf{x}_A \in Q\} \ge \mathrm{Var}\{f(\mathbf{x}_A) \mid \mathbf{x}_A \in Q\}$. Given $\mathbf{x}_A \in Q$, the coordinates are independent and uniform on their intervals, so $\mathrm{Var}(f \mid Q) = \sum_k \beta_k^2 h_{k,Q}^2/12$, where $h_{k,Q}$ is the length of the interval of coordinate k in Q. Averaging over cells with weights $\mathbb{P}(\mathbf{x}_A \in Q) = \prod_k h_{k,Q}$, and noting that the marginal probability of an interval on coordinate k equals its length,

$$\mathbb{E}(f-\bar{g})^2 \ge \sum_{k=1}^{s} \frac{\beta_k^2}{12} \sum_{l=1}^{L_k} h_{kl}\, h_{kl}^2 = \sum_{k=1}^{s} \frac{\beta_k^2}{12} \sum_{l=1}^{L_k} h_{kl}^3.$$

By the power-mean (or Jensen's) inequality, $\sum_{l=1}^{L} h_l^3 \ge L(L^{-1}\sum_l h_l)^3 = L^{-2}$, which gives the first bound.

Optimisation over the allocation of split points. A tree of depth d has at most $2^d - 1$ internal nodes, so $\sum_k K_k \le M(2^d - 1)$ and $\sum_k L_k \le S := M(2^d - 1) + s$. Relax the integer-valued L_k to positive real values and minimise $\phi(\mathbf{L}) = \sum_k \beta_k^2/(12L_k^2)$ subject to $\sum_k L_k \le S$. The function is convex and decreasing in each L_k, so the constraint is active; the Lagrange

condition $\beta_k^2/(6L_k^3) = \mu$ gives $L_k = C|\beta_k|^{2/3}$ with $C = S/\sum_k |\beta_k|^{2/3}$, and the minimum is $\sum_k |\beta_k|^{2/3}/(12C^2) = (\sum_k |\beta_k|^{2/3})^3/(12S^2)$. The minimum over integer values is at least as large as the minimum over real values. □

Proof of Theorem 6 Let $I_1, \dots, I_K$ be the intervals, of length $h = 1/K$, $N_k \geq 1$ the number of observations in I_k, $\bar{y}_k$ their mean response and $\bar{m}_k = \beta\bar{x}_k$ the mean of βx over these observations. Conditionally on $x_1, \dots, x_N$, $\bar{y}_k$ has mean $\bar{m}_k$ and variance σ^2/N_k, so for $x \in I_k$, $\mathbb{E}\{(\bar{y}_k - \beta x)^2 \mid x_1, \dots, x_N\} = \sigma^2/N_k + (\bar{m}_k - \beta x)^2$. The constant that minimises $c \mapsto \int_{I_k} (c - \beta x)^2 dx$ is the mean of βx over I_k, and the minimum is $\beta^2 h^3/12$ (the variance of a uniform variable on an interval of length βh, times h). Hence

$$\mathbb{E}\Big[\int_0^1 (\hat{g} - \beta x)^2 dx \;\Big|\; x_1, \dots, x_N\Big] \geq \sum_{k=1}^{K} \Big\{\frac{h\sigma^2}{N_k} + \frac{\beta^2 h^3}{12}\Big\} = \frac{\beta^2}{12K^2} + \frac{\sigma^2}{K}\sum_{k=1}^{K}\frac{1}{N_k}.$$

By the inequality between the arithmetic and harmonic means, $\sum_k 1/N_k \geq K^2/\sum_k N_k = K^2/N$, so the last term is at least $\sigma^2 K/N$. Let $A = \beta^2/12$ and $B = \sigma^2/N$. The function $K \mapsto A/K^2 + BK$ is convex on $(0, \infty)$ with minimiser $K = (2A/B)^{1/3}$ and minimum $A^{1/3}B^{2/3}(2^{-2/3} + 2^{1/3}) = 3(AB^2/4)^{1/3} = 3\{\beta^2\sigma^4/(48N^2)\}^{1/3}$. For the least-squares line, conditionally on the design the fitted value at x is unbiased with variance $\sigma^2\{N^{-1} + (x - \bar{x})^2/S_{xx}\}$, where $S_{xx} = \sum_i (x_i - \bar{x})^2$. Integrating over $[0, 1]$ gives $\sigma^2[N^{-1} + \{\frac{1}{12} + (\bar{x} - \frac{1}{2})^2\}/S_{xx}]$. Since $\bar{x} \to \frac{1}{2}$ and $S_{xx}/N \to \frac{1}{12}$ almost surely, this is $2\sigma^2 N^{-1}\{1 + o(1)\}$. □

Proof of Corollary 2 (i) With $\widehat{\mathbf{D}} = \mathbf{0}$ and $\boldsymbol{\beta}_z = \mathbf{0}$ the offsets vanish, no expectation step is performed, and only the fit of the trees on all measurements remains; with no linear component this is the XGBoost booster with a number of trees chosen on the folds. (ii) By Lemma 1 with $h \equiv 1$ and $\lambda = 0$ the leaf weights are the unshrunk mean residuals, which is the first-order least-squares boosting update of `gbm`; the outer loop is unchanged. (iii) With $f(\mathbf{x}) = a + \mathbf{x}^\top\boldsymbol{\beta}$ fitted by unpenalised least squares on the partial residuals, the fixed-effect update is the conditional maximiser of Q in $(a, \boldsymbol{\beta})$, and the expectation step and (8) are those of the expectation–conditional maximisation algorithm of the linear mixed model (Laird et al. 1987; Meng and Rubin 1993). □

Appendix B Tutorial

The method is implemented in the R package `MEXGB`, which uses `xgboost` (Chen and Guestrin 2016) for the trees and `glmnet` (Friedman et al. 2010) for the linear component. The main function is

```
MEXGB(X, Y, id, Z, time, family = "gaussian", iter = 100, ntree = 500,
      shrinkage = 0.05, interaction.depth = 2, n.minobsinnode = 2, lambda = 1,
      crossfit = 5, inner = 50, inner_tol = 1e-4, linear = "auto",
      lin_relax = TRUE, screen = "auto", ntree_select = "cv",
      delta = 0.001, patience = 5)
```

Here, `X` is the $N \times p$ predictor matrix, `Y` is the response vector, `Z` is the $N \times q$ random-effect design matrix, `id` contains subject identifiers, and `time` contains measurement times. The arguments correspond to the quantities defined in Section 3.3: `shrinkage`, `interaction.depth`, `n.minobsinnode` and `lambda` are ν, d, the minimum leaf size and the leaf penalty; `ntree` is $M_{\max}$ and `ntree_select = "cv"` selects $\hat{M}$ using out-of-fold

loss; `crossfit` is the number of folds V (`crossfit = 0` gives the in-sample expectation step); `inner` and `inner_tol` control the inner iterations; `linear`, `lin_relax` and `screen` control the linear component, its relaxation and the screening (the out-of-fold random-effect offsets (11) and the covariance check (13) are used by default); and `delta` and `patience` control the stopping rule. The following example uses earlier visits for training and later visits for testing. This chronological split differs from the random within-subject splits used in the reported benchmark.

```
devtools::install_github("rid4stat/MEXGB")
library(MEXGB)
set.seed(1)
d <- simLong(n = 50, p = 170, rel_p = 6, time_points = 20, linear = FALSE,
             random_sd_intercept = sqrt(0.5), random_sd_slope = sqrt(3),
             noise_sd = 0.5, rho_W = 0.6, rho_Z = 0.6, family = "gaussian")
X <- as.matrix(d[, grep("^Var", names(d))])
Z <- as.matrix(d[, c("RandomIntercept", "RandomSlope")])
tr <- d$time <= 0.7 # earlier visits for training
fit <- MEXGB(X[tr, ], d$Y[tr], id = d$id[tr], Z = Z[tr, ], time = d$time[tr],
           interaction.depth = 2, n.minobsinnode = 2, verbose = FALSE)
pred <- predict(fit, X[!tr, ], Z[!tr, ], id = d$id[!tr], time = d$time[!tr])
mean((d$Y[!tr] - pred)^2) # test MSE
fit$var_random_effects # estimated D
summary(fit, top = 10) # predictor importance
```

The fitted object contains the boosted ensemble and the linear component, the predicted random effects, the estimates of $\mathbf{D}$, σ^2 and $\boldsymbol{\beta}_z$, the out-of-fold criterion and the marginal log-likelihood over the outer iterations, and the variable-importance scores defined in (14). The `predict` method adds the estimated random-effect contribution for previously observed subjects and sets this contribution to zero for new subjects.

Appendix C Additional results

Table C1 Mean training time per split (seconds on one core) in the main scenarios.

	Linear			Nonlinear		
Method	6	170	2000	6	170	2000
MEXGB	6.1	118.8	912.4	65.9	111.1	561.6
MEGB	3.5	57.8	641.1	3.9	37.5	566.4
GPBoost	0.8	1.3	8.6	0.8	1.3	7.8
MERF	9.2	96.4	600.3	11.3	152.9	843.6
REEMforest	131.2	158.9	701.4	137.0	234.9	1115.1
XGBoost	0.2	2.7	28.7	0.2	2.4	29.1
GBM	0.1	1.3	16.8	0.1	1.2	17.8
RF	0.7	11.3	69.0	0.8	13.8	89.8
mboost	0.2	0.2	1.8	0.2	0.2	1.9
glmmLasso	0.8	1.7	–	0.8	0.6	–
LMM	0.1	–	–	0.1	–	–

Table C2 Mean training time per split (seconds on one core) on the real datasets.

Method	BodyWeight	Soybean	Sprockett
MEXGB	10.6	10.0	1653.3
MEGB	0.3	2.0	1965.4
GPBoost	0.7	0.7	9.4
MERF	0.5	5.5	2455.2
REEMforest	7.7	120.2	4062.8
XGBoost	0.1	0.1	42.4
GBM	0.0	0.0	30.8
RF	0.0	0.0	100.3
mboost	0.1	0.1	2.7

Table C3 Sprockett data: selection frequency (proportion of the 100 splits in which an ASV is among the six top-ranked predictors) of the eight ASVs selected most often by MEXGB.

ASV	MEXGB	MEGB	XGBoost	GBM	RF
5892 *Parabacteroides*	0.75	0.80	0.74	0.77	0.36
4085 *Veillonella*	0.71	0.37	0.81	0.71	0.99
5286 *Prevotella* 2	0.69	0.76	0.73	0.86	0.00
6272 *Ruminococcaceae* UCG-010	0.65	0.75	0.68	0.69	0.25
2208 *Ruminococcaceae*	0.58	0.41	0.62	0.61	0.00
3522 *Streptococcus anginosus*	0.46	0.00	0.50	0.41	0.62
1323 *Christensenellaceae* R-7	0.36	0.00	0.34	0.34	0.00
4327 *Collinsella aerofaciens*	0.27	0.00	0.39	0.00	1.00

References

Asadi F, Homayounfar R, Mehrali Y, et al (2026) Innovative statistical method for longitudinal and hierarchical data modeling: the gmexgboost method. BMC Medical Research Methodology 26:24

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics 37(4):1705–1732

Boucheron S, Lugosi G, Massart P (2013) Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, Oxford

Breiman L (2001) Random forests. Machine Learning 45(1):5–32

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

Capitaine L, Genuer R, Thiébaut R (2021) Random forests for high-dimensional longitudinal data. Statistical Methods in Medical Research 30(1):166–184. https://doi.org/10.1177/0962280220946080

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

Chernozhukov V, Chetverikov D, Demirer M, et al (2018) Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21(1):C1–C68

Cho SJ, Mueller S (2026) Mixed-effects XGBoost with group-aware permutation importance and cross-validation for multilevel cross-classified continuous outcomes. Psychometrika https://doi.org/10.1017/psy.2026.10108, published online 10 April 2026

Crowder MJ, Hand DJ (1990) Analysis of Repeated Measures. Chapman and Hall, London

Davidian M, Giltinan DM (1995) Nonlinear Models for Repeated Measurement Data. Chapman and Hall, London

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) 39(1):1–22

Diggle PJ, Heagerty P, Liang KY, et al (2002) Analysis of Longitudinal Data, 2nd edn. Oxford University Press, Oxford

Fitzmaurice GM, Laird NM, Ware JH (2011) Applied Longitudinal Analysis. John Wiley & Sons

Fokkema M, Smits N, Zeileis A, et al (2018) Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behavior Research Methods 50(5):2016–2034

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1):1–22

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of Statistics 29(5):1189–1232

Fu G, Fu H, Dai X, et al (2026) Longitudinal random forests for sparse and irregular response trajectories. arXiv preprint arXiv:260721817

Fu W, Simonoff JS (2015) Unbiased regression trees for longitudinal and clustered data. Computational Statistics & Data Analysis 88:53–74

Fuhr J, Papies D (2024) Double machine learning meets panel data – promises, pitfalls, and potential solutions. URL https://arxiv.org/abs/2409.01266, arXiv:2409.01266

Griesbach C, Säfken B, Waldmann E (2021) Gradient boosting for linear mixed models. The International Journal of Biostatistics 17(2):317–329

Groll A, Tutz G (2012) Regularization for generalized additive mixed models by likelihood-based boosting. Methods of Information in Medicine 51(2):168–177

Groll A, Tutz G (2014) Variable selection for generalized linear mixed models by l1-penalized estimation. Statistics and Computing 24(2):137–154

Hajjem A, Bellavance F, Larocque D (2011) Mixed effects regression trees for clustered data. Statistics & Probability Letters 81(4):451–459

Hajjem A, Bellavance F, Larocque D (2014) Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation 84(6):1313–1328

Hajjem A, Larocque D, Bellavance F (2017) Generalized mixed effects regression trees. Statistics & Probability Letters 126:114–118

Hastie T, Tibshirani R, Tibshirani R (2020) Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. Statistical Science 35(4):579–592

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31(2):423–447

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58(301):13–30

Hu J, Szymczak S (2023) A review on longitudinal data analysis with random forest. Briefings in Bioinformatics 24(2):bbad002. https://doi.org/10.1093/bib/bbad002, URL https://academic.oup.com/bib/article/24/2/bbad002/6991123

Knieper L, Hothorn T, Bergherr E, et al (2025) Gradient boosting for generalised additive mixed models. Statistics and Computing 35(4):1–20

Laird N, Lange N, Stram D (1987) Maximum likelihood computations with repeated measures: application of the EM algorithm. Journal of the American Statistical Association 82(397):97–105

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38(4):963–974

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, pp 4765–4774

Lundberg SM, Erion G, Chen H, et al (2020) From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2(1):56–67

McLachlan GJ, Krishnan T (2008) The EM Algorithm and Extensions, 2nd edn. Wiley, Hoboken, NJ

Meinshausen N (2007) Relaxed lasso. Computational Statistics & Data Analysis 52(1):374–393

Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80(2):267–278

Messer P, Schmid T (2024) Gradient boosting for hierarchical data in small area estimation. arXiv preprint arXiv:240604256

Ngufor C, Van Houten H, Caffo BS, et al (2019) Mixed effect machine learning: a framework for predicting longitudinal change in hemoglobin A1c. Journal of Biomedical Informatics 89:56–67

Olaniran OR, Olaniran SF, Allohibi J, et al (2025) Mixed effect gradient boosting for high-dimensional longitudinal data. Scientific Reports 15(1):30927

Pellagatti M, Masci C, Ieva F, et al (2021) Generalized mixed-effects random forest: a flexible approach to predict university student dropout. Statistical Analysis and Data Mining 14(3):241–257

Pinheiro JC, Bates DM (2000) Mixed-Effects Models in S and S-PLUS. Springer, New York

Prevett ML, Hui FKC, Tho ZY, et al (2025) Gradient boosted mixed models: Flexible estimation of mean and variance components for clustered data. arXiv preprint arXiv:251100217

Robinson GK (1991) That BLUP is a good thing: the estimation of random effects. Statistical Science 6(1):15–32

Schelldorfer J, Bühlmann P, van de Geer S (2011) Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. Scandinavian Journal of Statistics 38(2):197–214

Sela RJ, Simonoff JS (2012) Re-em trees: a data mining approach for longitudinal and clustered data. Machine Learning 86(2):169–207

Sigrist F (2022) Gaussian process boosting. Journal of Machine Learning Research 23(232):1–46

Sigrist F (2023) Latent Gaussian model boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(2):1894–1905

Speiser JL, Wolf BJ, Chung D, et al (2019) BiMM forest: a random forest method for modeling clustered and longitudinal binary outcomes. Chemometrics and Intelligent Laboratory Systems 185:122–134

Sprockett DD, Martin M, Costello EK, et al (2020) Microbiota assembly, structure, and dynamics among tsimane horticulturalists of the bolivian amazon. Nature Communications 11(1):3772

Tan YV, Flannagan CAC, Elliott MR (2018) Predicting human-driving behavior to help driverless vehicles drive: random intercept Bayesian additive regression trees. Statistics and Its Interface 11(4):557–572

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58(1):267–288

Toivonen J, Kangas A, Pitkänen TP, et al (2026) Combining tree-boosting and mixed effects models improves the performance of remote-sensing based forest age predictions. Forestry: An International Journal of Forest Research 99(3):cpaf083. https://doi.org/10.1093/forestry/cpaf083

Tutz G, Groll A (2010) Generalized linear mixed models based on boosting. In: Kneib T, Tutz G (eds) Statistical Modelling and Regression Structures. Physica-Verlag, Heidelberg, p 197–215

van der Vaart AW (1998) Asymptotic Statistics. Cambridge University Press, Cambridge

Verbeke G, Molenberghs G (2000) Linear Mixed Models for Longitudinal Data. Springer, New York

Wu CFJ (1983) On the convergence properties of the EM algorithm. The Annals of Statistics 11(1):95–103

Yang L, Sun E, Wu TT (2024) splmm: Simultaneous Penalized Linear Mixed Effects Models. URL https://cran.r-project.org/web/packages/splmm/splmm.pdf, r package version 1.2.0

Zhang T, Yu B (2005) Boosting with early stopping: convergence and consistency. The Annals of Statistics 33(4):1538–1579