Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Fixed Effects Testing in High-Dimensional Linear Mixed Models

Jelena Bradic , Gerda Claeskens & Thomas Gueuning

To cite this article: Jelena Bradic , Gerda Claeskens & Thomas Gueuning (2020) Fixed Effects Testing in High-Dimensional Linear Mixed Models, Journal of the American Statistical Association, 115:532, 1835-1850, DOI: 10.1080/01621459.2019.1660172

To link to this article: https://doi.org/10.1080/01621459.2019.1660172

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 1835–1850: Theory and Methods
https://doi.org/10.1080/01621459.2019.1660172

Taylor & Francis
Taylor & Francis Group

Fixed Effects Testing in High-Dimensional Linear Mixed Models

Jelena Bradic[a], Gerda Claeskens[b], and Thomas Gueuning[b]

[a]Department of Mathematics, University of California San Diego, La Jolla, CA; [b]ORStat and Leuven Statistics Research Center, KU Leuven, Belgium

ABSTRACT

Many scientific and engineering challenges—ranging from pharmacokinetic drug dosage allocation and personalized medicine to marketing mix (4Ps) recommendations—require an understanding of the unobserved heterogeneity to develop the best decision making-processes. In this article, we develop a hypothesis test and the corresponding *p*-value for testing for the significance of the homogeneous structure in linear mixed models. A robust matching moment construction is used for creating a test that adapts to the size of the model sparsity. When unobserved heterogeneity at a cluster level is constant, we show that our test is both consistent and unbiased even when the dimension of the model is extremely high. Our theoretical results rely on a new family of adaptive sparse estimators of the fixed effects that do not require consistent estimation of the random effects. Moreover, our inference results do not require consistent model selection. We showcase that moment matching can be extended to nonlinear mixed effects models and to generalized linear mixed effects models. In numerical and real data experiments, we find that the developed method is extremely accurate, that it adapts to the size of the underlying model and is decidedly powerful in the presence of irrelevant covariates. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2017
Accepted July 2019

KEYWORDS
Misspecification; Penalization; *p*-Values; Random effects; Robustness

1. Introduction

In many applications, we want to use data to draw inferences about the common underlying effect of a treatment. Examples include medical studies about the computation of effect sizes for assessing the clinical importance of covariates on pharmacokinetic or pharmacodynamic responses, and to the study of drug–drug interactions or studies that quantify the effects of different advertising mediums that take into account other variables such as pricing, distribution points, and competitor tactics, for example, commonly used by technology firms to optimize budget over these different mediums. Historically, most datasets have been too small to meaningfully explore unobserved heterogeneity beyond dividing the sample into independent subgroups. Recently, however, there has been an explosion of empirical settings where it is potentially feasible to gather large-scale observations and therefore better customize estimates across both population and individuals.

An impediment to exploring unobserved heterogeneous effects is the fear that researchers will iteratively search for subgroups with high treatment levels, and then report only the results for subgroups with extreme effects, thus exploring and utilizing heterogeneity that may be purely spurious. Moreover, procedural restrictions have often been used to control for the unobserved randomness. However, such procedural restrictions can make it difficult to encompass strong but unexpected heterogeneity which naturally occurs in practice. In this article, we seek to address this challenge by developing a method for hypothesis testing of fixed effects, while allowing models to have heterogeneous and random components, that yields valid asymptotic confidence intervals and *p*-values for the true underlying fixed effect.

Classical approaches to test fixed effects in the presence of random effects include Breslow and Clayton (1993), Kenward and Roger (1997), and Crainiceanu and Ruppert (2004); see, for example, Verbeke and Molenberghs (2009). These methods perform well in applications with a small number of covariates, but quickly break down as the number of covariates increases. In this article, we explore the use of doubly robust ideas from the literature to improve the performance of these classical methods in the presence of an exploding number of covariates. We develop a family of moment matching tests, which allows for flexible modeling of interactions in high dimensions by allowing models that are not entirely sparse (not even approximately). The developed moments are related to Neyman orthogonalization principles in that they are robust to model misspecifications and estimation of nuisance parameters; however, the tests differ in that they are also robust to the misspecification (and/or misestimation) of random effects in the case of linear mixed models—this is especially important in environments with many covariates with possible complex interactions and the presence of unobserved model heterogeneity.

Despite their widespread success in estimation and inference in high-dimensional linear models, there are important hurdles that need to be cleared before penalized estimators are directly useful in linear mixed models. Ideally, an estimator of complete variability in the model needs to be constructed, so that a researcher can use it to test hypotheses and establish sampling distributions. However, in the case of mixed models, estimation

CONTACT Jelena Bradic *jbradic@ucsd.edu* Department of Mathematics and Halicioglu Data Science Institute, University of California at San Diego, 9500 Gilman Drive 0120, La Jolla, CA.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

of random effects is typically more difficult than that of fixed effects one is interested in. Developing inferential tools that do not necessarily rely on good quality estimates of random effects would therefore be extremely useful. Yet, such procedures have been largely left undeveloped, even in the standard contexts.

This article addresses these limitations, develops a robust method for fixed effect testing that allows for a tractable asymptotic theory and valid statistical inference even when the estimation of random effects is not consistent. Our proposed test is composed of an estimator of the correlation between the model error under the null hypothesis and the error of carefully constructed feature projections.

In the interest of generality, we begin our theoretical analysis by developing the desired consistency and asymptotic normality results in the context of high-dimensional linear mixed models. We prove these results for a particular variant of mixed models that uses Gaussian random effects with unknown cluster variances while allowing the number of fixed effects to be much larger than the sample size. However, the results do not rely on Gaussianity or quality of estimators of the unknown variance components. This property is achieved by the doubly robust construction of the test statistic, as well as new high-dimensional estimators of the fixed effects. We also show that such robustly motivated estimator is consistent in l_1 norm whenever the actual model is sparse. Our proof is built on within cluster analysis and a martingale representation of the test statistic between the clusters. We show that the consistency of the test adapts to the quality of estimation of the fixed effects by successfully leveraging correlation properties among the features. Given these general results, we next show that our ideas extend from the linear setting to the nonlinear setting as well as settings with unknown cluster correlation. We also illustrate that the proposed fixed effect estimator can be utilized for estimation of the unknown variance components.

Although our main focus in this article is the inference on fixed or common effects, we note that there are a variety of important applications of the p-value construction in a pure variable selection context. Ryzhov, Han, and Bradic (2016) seek to improve the construction of marketing campaigns for nonprofit organizations by detecting which marketing strategy yielded a better donor retention (in terms of continuous influx of small to medium donations). Donors are grouped naturally by the observed frequency of donations thus yielding a natural longitudinal structure in the observations. Here we need rigorous predictions for the probability that a donor would continue its activity if a specific marketing strategy was selected as the most beneficial for the observed data. Our results would be one of the first to develop rigorous variable selection that enables the use of large-scale data for this purpose.

1.1. Related Work

There has been a longstanding understanding in the high-dimensional statistics literature that prediction methods based on regularization (e.g., Tibshirani 1996; Fan and Li 2001) perform well outside of the class of linear models: if the goal is prediction, then we should define a proper likelihood function and the method will be considered as good as long as the likelihood function is convex (see Bühlmann and Van De Geer 2011 for more details). However, good performance in prediction does not necessarily translate into good performance for estimation or inference about model parameters. In fact, it can often be quite poor. In a Neyman orthogonalization framework we use to formalize our inferential results, we show that the testing of significance of fixed effect can be done asymptotically exactly regardless of the error in estimation of the random effects or the additional nuisance parameters of the model. Thus, when evaluating estimators of the fixed effects, asymptotic theory plays a much more important role than in the standard prediction context.

From a technical point of view, the main contribution of this article is an asymptotic normality theory enabling statistical inference in the context of high-dimensional linear mixed models while simultaneously allowing misspecification in both the model and the random effects. Recent results by Zhang and Zhang (2014), Van de Geer et al. (2014), Javanmard and Montanari (2014a), Cai and Guo (2017), Belloni, Chernozhukov, and Kato (2014), and others have established asymptotic properties of tests in a particular variant of the sparse high-dimensional linear models (Ren et al. 2015; Jankova and Van De Geer 2015; Ning and Liu 2017; Athey, Imbens, and Wager 2016). To our knowledge, however, we provide the first set of conditions under which tests are both asymptotically unbiased and Gaussian for the linear mixed models, thus allowing for classical statistical inference; the estimator of the fixed effects used to achieve asymptotic normality proposed in this article is also new. We review the existing theoretical literature on linear mixed models in more detail in Section 3.1.

A small but growing literature, including Belloni et al. (2015) and Chernozhukov, Hansen, and Spindler (2015), has considered the use of Neyman orthogonalization for the purposes of high-dimensional inference. These articles use the Neyman orthogonalization method of Neyman (1959) together with sample splitting and de-biasing (e.g., Chernozhukov et al. 2016), and report confidence intervals or p-values resulting from significance testing, obtained by Belloni et al. (2017). A limitation of the existing work is in that it cannot allow for the unobserved heterogeneity in the model (e.g., like the presence of random effects).

We view our contribution as complementary to this literature, by showing that Neyman orthogonalization need not only be viewed as successful in linear models (or generalized linear models), and can instead be modified and used for rigorous asymptotic analysis in high-dimensional model with random effects. We succeed in showing that our construction allows one more degree of robustness—the random effects can be estimated rather poorly without changing the asymptotic null distribution of the test. Even in low dimensions advancements of this type would be of considerable interests as there are no fully reliable ways to identify the best covariance model; as many studies have revealed biased inference for the fixed effects with an underspecified covariance structure; the problem is only multiplied when there is a growing number of fixed effects in the model. The methodological and theoretical tools developed here are useful beyond the specific class of algorithms studied in our article. In particular, our tools allow for a fairly direct analysis

of variants of the Bayesian hierarchical linear regression models (e.g., Quintana et al. 2016).

Finally, we note a growing literature on estimating fixed effects in the presence of unobserved heterogeneity using different regularization methods. Schelldorfer, Buhlmann, and van de Geer (2011), Groll and Tutz (2014), and Hui, Müller, and Welsh (2017) developed lasso-like methods for estimation of fixed effects in a sparse high-dimensional linear mixed model setting. Nonconvex methods were investigated by Wang, Zhou, and Qu (2012) and Ghosh and Thoresen (2016), among others. Fan and Li (2012) and others discussed variable selection procedures for the fixed or random effects that enable off-the-shelf methods to be used for optimal estimation (see a review article Müller, Scealy, and Welsh (2013) for more details). Additionally, Bonnet, Gassiat, and Lévy-Leduc (2015) estimated heritability in sparse linear mixed models, relying on a special asymptotic scenario where the sizes of the clusters are proportional to the number of observations. However, none of the approaches above provides guarantees about honest p-values or confidence interval constructions.

2. Inference in Linear Mixed Models

We consider a linear mixed model in which observations are grouped. Suppose that we have a sample of N subjects. Let $i = 1,\dots,N$ be the grouping index and n_i the number of observations in group i. The total number of observations is denoted by $n = \sum_{i=1}^{N} n_i$. For the ith subject we represent the model as follows

$$\mathbf{y}_i = \boldsymbol{X}_i\boldsymbol{\gamma}^* + \boldsymbol{Z}_i\beta^* + \mathbf{W}_i\boldsymbol{b}_i + \epsilon_i, \tag{1}$$

where $\mathbf{y}_i = (y_{i1},\dots,y_{in_i})^\top \in \mathbb{R}^{n_i}$ collects the response variable of the ith subject, $\boldsymbol{X}_i \in \mathbb{R}^{n_i\times(p-1)}$ with $\boldsymbol{X}_i = (x_{i1}^\top, x_{i2}^\top,\dots,x_{in_i}^\top)^\top$ the design matrix of the fixed effects where $x_{ij} \in \mathbb{R}^{p-1}$ for $j \in \{1,\dots,n_i\}$. Similarly we denote with $\boldsymbol{Z}_i = (z_{i1}, z_{i2},\dots,z_{in_i})^\top$ the design vector of the fixed effects which we are interested in, with $z_{ij} \in \mathbb{R}$. The vector of population specific fixed effect coefficients is split into a $(p-1)$-dimensional vector $\boldsymbol{\gamma}^*$ and a univariate β^*. The subject specific random effects are defined through a q dimensional vector $\boldsymbol{b}_i = (b_{i1},\dots,b_{iq})^\top \in \mathbb{R}^q$ for which we assume $\boldsymbol{b}_i \sim \mathcal{N}_q(0,\boldsymbol{\psi})$ for a bounded q. Here, $\boldsymbol{\psi}$ is the unknown covariance matrix of the random effects which can be correlated with each other. The corresponding deterministic design matrix for group i is denoted with $\boldsymbol{W}_i \in \mathbb{R}^{n_i\times q}$. In addition, ϵ_i is the random error vector with components iid with mean zero and an unknown variance $0 < \sigma_\epsilon^2 < \infty$. We would like to remark that the independence assumption can be generalized to a positive definite variance covariance structure. This generalization still fits into the theoretical framework presented in Section 3. Nonetheless, for the sake of notational simplicity, we restrict to the case of group independent errors.

The goal of this article is to develop a test which is able to detect whether β^* is equal to β_0 or not, that is,

$$H_0 : \beta^* = \beta_0 \text{ versus } H_1 : \beta^* \neq \beta_0, \tag{2}$$

for some given value $\beta_0 \in \mathbb{R}$. We study the asymptotics with $N \to \infty$ and assume that n_i is bounded in N, while allowing a large number of fixed effects in that $p \gg n$, that is, $\log p = o(\sqrt{n})$. The main difficulty is that we can only ever test the fixed effects if the random effects are either known or estimated well. However, in practice this is never achievable and we cannot directly test for the fixed effects using existing tools and techniques. In general, we cannot estimate the variance components consistently simply from the observed data without further restrictions on the data generating distribution. A standard way to make progress is to assume models selection consistency of the fixed effects, that is, that estimated fixed effects are correctly selected. The motivation behind this assumption is that it enables direct dimensionality reduction as it effectively implies correct selection of the true fixed effects; thus, imposing restrictive minimal signal strength assumptions together with a irrepresentable condition. In this article, we take a more indirect approach: we show that, under simple assumptions, our approach can use moment conditions to achieve consistency of a test without needing to explicitly estimate the variance component.

Let vectors $\boldsymbol{Y}$, $\boldsymbol{b}$, and $\boldsymbol{\epsilon}$, and matrices $\boldsymbol{X}$, $\boldsymbol{Z}$ be obtained by stacking vectors $\mathbf{y}_i$, $\boldsymbol{b}_i$, and ϵ_i and matrices $\boldsymbol{X}_i, \boldsymbol{Z}_i$, respectively, underneath each other, and let $\boldsymbol{\Psi} = \operatorname{diag}(\boldsymbol{\psi},...,\boldsymbol{\psi}) \in \mathbb{R}^{qN\times qN}$ so that $\boldsymbol{Wb} \sim \mathcal{N}_n(0, \boldsymbol{W\Psi W}^\top)$ with a block-diagonal matrix $\boldsymbol{W} = \operatorname{diag}(\boldsymbol{W}_1,\dots,\boldsymbol{W}_N) \in \mathbb{R}^{n\times qN}$. In particular, $\boldsymbol{b} = (\boldsymbol{b}_1^\top,\dots,\boldsymbol{b}_N^\top)^\top \in \mathbb{R}^{qN}$ is the random effect vector and $\boldsymbol{Y}|(\boldsymbol{X},\boldsymbol{Z})$ has mean $\boldsymbol{X\gamma}^* + \boldsymbol{Z\beta}^*$ and variance $\sigma_\epsilon^2\boldsymbol{I}_n + \boldsymbol{W\Psi W}^\top$. We further standardize the design matrix $\boldsymbol{X}$ such that each column has the same norm. The linear mixed effects model (1) can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{X\gamma}^* + \boldsymbol{Z}\beta^* + \boldsymbol{Wb} + \boldsymbol{\epsilon}. \tag{3}$$

The n components of the noise vector $\boldsymbol{\epsilon}$ are iid with mean zero and variance $0 < \sigma_\epsilon^2 < \infty$. We assume that $\boldsymbol{\epsilon}$, $\boldsymbol{b}$, and $\boldsymbol{W}$ are mutually independent and that $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{X},\boldsymbol{Z}$. Observe that we do not require the error to have Gaussian distribution (see Schelldorfer, Buhlmann, and van de Geer (2011) and Fan and Li (2012), where Gaussianity was explicitly assumed); if we restrict our attention to Gaussian error then the independence above can be replaced with uncorrelatedness with no significant changes in the proofs. Note that $\boldsymbol{\gamma}^*$ and β^* can be seen as being derived from an original p-dimensional vector. The notation (3) emphasizes the fact that β^* is the component of this original vector on which we want to perform hypothesis testing.

2.1. From the Parametric Null to the Matching Moments Condition

At a high level, Wald and score tests can be thought of as two contrasting methods with likelihood based constructions. Given a particular score function (typically a first derivative of the likelihood function), a classical method such as Rao's test, performs estimation only under the null hypothesis and forms a test based on the first moment condition, that is the expectation of the score should be close to zero. In contrast, orthogonalization methods seek to find appropriate directions that are close to the score, where the closeness is defined with respect to a projection of a certain kind. The advantage of orthogonalization is that hypothesis testing of a parameter of primary interest can be done rather efficiently in the presence of the nuisance

parameters. This is achieved by constructing a class of functions (dependent on the null hypothesis) that is orthogonal to the scores of the nuisance parameters. Neyman then considers a particular construction of regressing the scores of the parameter of interest onto the scores of the nuisance parameters. Then the test, formulated by exploring this orthogonality, can potentially lead a to a substantial increase in power.

In this section, we seek orthogonalization principles that are adapted to the presence of random effects in a linear regression model. Suppose first that we observe independent samples $(\mathbf{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ and want to build a orthogonalization criterion suitable for testing (2) in a model (3). We start by splitting the feature space until we have partitioned it into $\boldsymbol{Z}_i$, a set of covariates in the null, and $\boldsymbol{X}_i$, associated with the nuisance parameters. Then, given observations $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ only, we evaluate the regression of $\boldsymbol{Z}_i$ onto the remaining covariates $\boldsymbol{X}_i$ by setting

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{U}, \tag{4}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^{p-1}$ and with components of $\boldsymbol{U}$ are iid with mean zero and unknown variance $0 < \sigma_u^2 < \infty$. Heuristically, this strategy is well-motivated if we believe the covariates of the fixed design have a Gaussian distribution and that the rows are roughly identically distributed. There are possibly several other procedures on how to best design this regression; but for simplicity we consider a linear case—other more elaborate cases follow from our methodology with easy extensions of the proofs.

We assume that $\boldsymbol{U}$ is independent of $\boldsymbol{X}$, $\boldsymbol{\epsilon}$, and $\boldsymbol{b}$. However, uncorrelatedness can replace the independence with little changes in the proof whenever the errors in the model take Gaussian structure. Note that the coefficients $\boldsymbol{\theta}^*$ are functions of the variances and covariances of the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, that is, $\boldsymbol{\theta}^*$ and of σ_u^2 depend on $\mathcal{X} = (\boldsymbol{X}, \boldsymbol{Z})$. They can be explicitly computed if the rows of $\mathcal{X}$ are iid with normal distribution $\mathcal{N}_p(0, \boldsymbol{\Sigma})$. In that case, denoting by j the column of $\mathcal{X}$ corresponding to $\boldsymbol{Z}$, it follows from Anderson (1984) that the conditional density $f(Z_i|\boldsymbol{X}_i)$ is a $(p-1)$-variate normal with mean $\boldsymbol{X}_i^\top \boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}$ and variance $\sigma_u^2 = \boldsymbol{\Sigma}_{j,j} - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}$, where $\boldsymbol{\Sigma}_{-j,-j} \in \mathbb{R}^{(p-1)\times(p-1)}$ is obtained by removing the jth row and column of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_{-j,j} \in \mathbb{R}^{p-1}$ is the jth column of $\boldsymbol{\Sigma}$ without its jth component. It follows that $\boldsymbol{\theta}^* = \boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}$. Table 1 gives the exact values of $\boldsymbol{\theta}^*$ and σ_u^2 for the setting considered in our numerical work. More details are provided in the Appendix.

Finally, given regression (3) together with (4) our procedure for testing (2), generates the pseudo response

$$\boldsymbol{V} = \boldsymbol{Y} - \boldsymbol{Z}\beta_0 \tag{5}$$

Table 1. Exact values of $\boldsymbol{\theta}^*$ and σ_u^2 for the settings considered in our simulation study.

Design	$\boldsymbol{\theta}^*$	Particular case
Toeplitz	Sparse	$\rho = -0.5 \Rightarrow \begin{cases} \boldsymbol{\theta}^* = (0, \ldots, 0, -0.4, -0.4, 0, \ldots, 0) \\ \sigma_u^2 = 0.6 \end{cases}$
Equi-correlated	Dense	$\rho = 0.8 \Rightarrow \begin{cases} \boldsymbol{\theta}^* = (0.002, \ldots, 0.002) \\ \sigma_u^2 = 0.2 \end{cases}$

NOTE: Note that for the Toeplitz case, if $\boldsymbol{Z}$ the first (resp., the last) column of $\mathcal{X}$ is tested then $\boldsymbol{\theta}^*$ is slightly different; only its first (resp., last) component is nonzero, with value ρ. Furthermore in that case, $\sigma_u^2 = 1 - \rho^2$.

and observes that under the null hypothesis $\boldsymbol{V} - \boldsymbol{X}\boldsymbol{\gamma}^* = \boldsymbol{W}\boldsymbol{b} + \boldsymbol{\epsilon}$ and $\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\theta}^* = \boldsymbol{U}$ are uncorrelated and are such that

$$\mathbb{E}[n^{-1}(\boldsymbol{V} - \boldsymbol{X}\boldsymbol{\gamma}^*)^\top \boldsymbol{A}(\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\theta}^*)] = 0 \tag{6}$$

for a $n \times n$ positive definite matrix $\boldsymbol{A}$. This moment matching equation can be seen as a specific orthogonality condition where $\boldsymbol{Z}_i$ are treated as confounders in the model on $\mathbf{V}$. Conversely, we observe that under the alternative hypothesis the above two terms are correlated since $\boldsymbol{V} - \boldsymbol{X}\boldsymbol{\gamma}^* = \boldsymbol{Z}(\beta^* - \beta_0) + \boldsymbol{W}\boldsymbol{b} + \boldsymbol{\epsilon}$; we compute that

$$\mathbb{E}[n^{-1}(\boldsymbol{V} - \boldsymbol{X}\boldsymbol{\gamma}^*)^\top \boldsymbol{A}(\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\theta}^*)] = \sigma_u^2 n^{-1}\text{trace}(\boldsymbol{A})(\beta^* - \beta_0).$$

The advantage of the above orthogonalization, (6), is that the form of a moment above resembles double-robust constructions where only one of the two residuals, $\boldsymbol{\epsilon}$ or $\boldsymbol{U}$, needs to be estimated well enough. Double robustness is to be understood in the sense that either model (4) can be misspecified but not both. Due to the assumed independence between $\boldsymbol{U}$ and $\boldsymbol{X}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{b}$, a nonzero correlation between $\boldsymbol{V} - \boldsymbol{X}\boldsymbol{\gamma}^*$ and $\boldsymbol{U} = \boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\theta}^*$ does not occur under the null hypothesis. Even if $\boldsymbol{\epsilon}$ contains further random effects structures, as long as they are independent/uncorrelated of the fixed effects, the correlation under the null should stay zero. Of course gross misspecification (of the likes of using a linear model when the model is highly nonlinear) is not considered. Misspecified is meant in light of the sparsity assumptions in each of the two models.

We will see that for a particular choice of the matrix $\boldsymbol{A}$ the above observation leads to optimal inference for the fixed effects without requiring any knowledge or even consistent estimation of the random effects. Such approach is of interest on its own right (even for low-dimensional problems) and would be extremely beneficial for practical purposes in high-dimensional longitudinal studies where estimation of random effects is particularly difficult.

2.2. Estimation of the Unknowns

In our discussion so far, we have emphasized the flexible nature of our methods: for a wide variety of the structure of $\boldsymbol{\gamma}^*$, distributions of $\boldsymbol{\epsilon}$ and of the random effects, $\boldsymbol{\Psi}$, our methods can be tailored, we achieve both consistency and asymptotic normality, provided the sample size n scales at an appropriate rate. Our results do, however, require the features corresponding to the fixed-effects are not correlated extremely highly: features are sparsely correlated if the precision matrix is row-sparse, that is, each row has small number of nonzero elements. We discuss a new class of estimators for both $\boldsymbol{\gamma}^*$ and $\boldsymbol{\theta}^*$ that satisfy this condition and are adaptive to the structure of the random effects.

Our first algorithm, which we call a doubly robust estimator, achieves adaptivity in estimation of $\boldsymbol{\gamma}^*$ regardless of its structure and the consistent estimation of the random effects. It achieves honest estimation by incorporating a pseudo response vector $\boldsymbol{V}$, (5), and a proxy correlation matrix $\tilde{\boldsymbol{P}}$ directly into its construction. To motivate our estimation procedure, let us observe that if $\boldsymbol{\epsilon}$ would be normally distributed, the underlying joint density

of $(\boldsymbol{V}, \boldsymbol{b})$ would be

$$f(\boldsymbol{V}, \boldsymbol{b}) = f(\boldsymbol{V}|\boldsymbol{b})f(\boldsymbol{b}) = (2\pi)^{-(n+qN)/2}\sigma_\epsilon^{-n}|\boldsymbol{\Psi}|^{-1/2} \times \exp\left\{-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{V}-\boldsymbol{X\gamma}-\boldsymbol{Wb})^\top(\boldsymbol{V}-\boldsymbol{X\gamma}-\boldsymbol{Wb}) - \frac{1}{2}\boldsymbol{b}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{b}\right\}. \quad (7)$$

We treat $f(\boldsymbol{V}, \boldsymbol{b})$ as a quasi-likelihood function; our method does not require normality of $\boldsymbol{\varepsilon}$. For a given $\boldsymbol{\gamma}$, the maximum likelihood estimator of $\boldsymbol{b}$ is then

$$\hat{\boldsymbol{b}}(\boldsymbol{\gamma}) = \left(\boldsymbol{W}^\top\boldsymbol{W} + \sigma_\epsilon^2\boldsymbol{\Psi}^{-1}\right)^{-1}\boldsymbol{W}^\top(\boldsymbol{V}-\boldsymbol{X\gamma}).$$

We can plug-in this estimator into (7) to construct a profile likelihood for $\boldsymbol{\gamma}$. We define $\boldsymbol{E} = \boldsymbol{W}^\top\boldsymbol{W} + \sigma_\epsilon^2\boldsymbol{\Psi}^{-1}$ and $\boldsymbol{P} = (\boldsymbol{I}_n - \boldsymbol{WE}^{-1}\boldsymbol{W}^\top)(\boldsymbol{I}_n - \boldsymbol{WE}^{-1}\boldsymbol{W}^\top) + \sigma_\epsilon^2\boldsymbol{WE}^{-1}\boldsymbol{\Psi}^{-1}\boldsymbol{E}^{-1}\boldsymbol{W}^\top$ and, dropping the constants, we obtain the following profile log-likelihood

$$\mathcal{L}(\boldsymbol{\gamma}, \hat{\boldsymbol{b}}(\boldsymbol{\gamma})) = -\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{V}-\boldsymbol{X\gamma})^\top\boldsymbol{P}(\boldsymbol{V}-\boldsymbol{X\gamma}). \quad (8)$$

By Lemma B.1 in the supplementary materials, $\boldsymbol{P} = (\boldsymbol{I}_n + \sigma_\epsilon^{-2}\boldsymbol{W\Psi W}^\top)^{-1}$. Note that if $\boldsymbol{\epsilon}$ is normally distributed then $\boldsymbol{Wb} + \boldsymbol{\epsilon} \sim \mathcal{N}_n(0, \sigma_\epsilon^2\boldsymbol{P}^{-1})$. However, observe that the above profile log-likelihood is a nonconvex function of all of the unknown parameters, $\boldsymbol{\gamma}, \boldsymbol{\Psi}, \sigma_\epsilon$. Moreover, in the presence of high-dimensional fixed effects, the profile log-likelihood has too many stationary points.

Following the idea of Fan and Li (2012), we replace the unknown variance matrix $\sigma_\epsilon^{-2}\boldsymbol{\Psi}$ by a known fixed symmetric matrix $\boldsymbol{M}$ which serves as a proxy. We discuss in the next section how to choose $\boldsymbol{M}$; our theory allows for many choices of $\boldsymbol{M}$. We define proxy versions of $\boldsymbol{E}$, $\boldsymbol{P}$, and $\mathcal{L}$ by replacing $\sigma_\epsilon^{-2}\boldsymbol{\Psi}$ by $\boldsymbol{M}$; we define $\tilde{\boldsymbol{E}} = \boldsymbol{W}^\top\boldsymbol{W} + \boldsymbol{M}^{-1}$ and $\tilde{\boldsymbol{P}} = (\boldsymbol{I}_n + \boldsymbol{WMW}^\top)^{-1}$ and obtain the following proxy log-likelihood

$$\tilde{\mathcal{L}}(\boldsymbol{\gamma}, \hat{\boldsymbol{b}}(\boldsymbol{\gamma})) = -\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{V}-\boldsymbol{X\gamma})^\top\tilde{\boldsymbol{P}}(\boldsymbol{V}-\boldsymbol{X\gamma}). \quad (9)$$

Observe that the proxy log-likelihood does not have a good quality approximation of the true $\mathcal{L}$. Instead it serves as a proxy to give inspiration for a new estimator of the unknown fixed effects $\boldsymbol{\gamma}^*$. Provided that the l_1 penalization of the above proxy log-likelihood heavily depends on the correct model specification and the form of the likelihood function, we take on a different perspective and design a new estimator that is inspired by the Dantzig selector. We define $\hat{\boldsymbol{\gamma}}$ as the solution to the following optimization problem

$$\hat{\boldsymbol{\gamma}} = \begin{array}{ll}\operatorname{argmin}_{\gamma\in\mathbb{R}^{p-1}} & \|\boldsymbol{\gamma}\|_1 \\ \text{subject to} & \left\|n^{-1}\boldsymbol{X}^\top\tilde{\boldsymbol{P}}(\boldsymbol{V}-\boldsymbol{X\gamma})\right\|_\infty \le \eta_\gamma \\ & n^{-1}\boldsymbol{V}^\top\tilde{\boldsymbol{P}}(\boldsymbol{V}-\boldsymbol{X\gamma}) \ge \bar{\eta}_\gamma \\ & \left\|\tilde{\boldsymbol{P}}(\boldsymbol{V}-\boldsymbol{X\gamma})\right\|_\infty \le \mu_\gamma\end{array} \quad (10)$$

with $\eta_\gamma \asymp (\log n)\sqrt{n^{-1}\log p}$, $\mu_\gamma \asymp \sqrt{\log n}$ and $0 < \bar{\eta}_\gamma < n^{-1}\operatorname{trace}(\sigma_\epsilon^2\tilde{\boldsymbol{P}}\boldsymbol{P}^{-1})$ suitably chosen tuning parameters. The l_1 norm induces sparse solutions whereas the constraints ensure good theoretical properties. The above constraints arise from the score vector of the profile log-likelihood function $\tilde{\mathcal{L}}$; the first ensures that the score vector is minimized and the second ensures that the variance of the residuals is properly guessed. However, observe that we do not assume that the profile log-likelihood is correctly specified, and for that matter a last constraint is needed to guarantee small size of the estimated residuals in the model

$$\tilde{\boldsymbol{P}}^{1/2}\boldsymbol{V} = \tilde{\boldsymbol{P}}^{1/2}\boldsymbol{X\gamma} + \boldsymbol{e}, \quad (11)$$

where $\operatorname{var}(\boldsymbol{e}) = \sigma_\epsilon^2\tilde{\boldsymbol{P}}\boldsymbol{P}^{-1}$. The tuning parameter $\bar{\eta}_\gamma$ provides a proxy for the unknown variance of the error in the model above. Hence, the size of the residuals is constrained in a similar way as robust loss functions are truncated to down-weight the effect of overly large outliers.

Remark 1. The estimator $\hat{\boldsymbol{\gamma}}$ defined in (10) is new and of potential interest in its own right. We show that under suitable conditions, $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 = O_P(\eta_\gamma\|\boldsymbol{\gamma}^*\|_0)$ meaning that this estimator is consistent if $\|\boldsymbol{\gamma}^*\|_0 = o(\sqrt{n}/\log(p)/\log(n))$. It is worth pointing that this result does not impose any restrictions on the minimum signal strength or that the matrix $\tilde{\boldsymbol{P}}$ correctly specifies variance parameters of the random effects and in that sense is honest and robust. Beside providing a reliable estimator, the optimization problem (10) is fast since it can be reformulated as a linear program. Existing regularized schemes, such as the one introduced by Schelldorfer, Buhlmann, and van de Geer (2011) for example, do not provide such quick implementation.

The estimation of $\boldsymbol{\theta}^*$ is based on model (4). Unlike linear models, where node-wise lasso is sufficient to estimate feature correlation, for the case of linear mixed models, we observe that the matrix $\tilde{\boldsymbol{P}}$ should contribute to the estimation procedure—we see from (11) that the covariates are premultiplied with $\tilde{\boldsymbol{P}}^{1/2}$. Namely, the presence of random effects induces a larger than usual dependence in the design matrix of the fixed effects. We incorporate such requirement in a new estimator defined below.

$$\hat{\boldsymbol{\theta}} = \begin{array}{ll}\operatorname{argmin}_{\theta\in\mathbb{R}^{p-1}} & \|\boldsymbol{\theta}\|_1 \\ \text{subject to} & \left\|n^{-1}\boldsymbol{X}^\top(\boldsymbol{Z}-\boldsymbol{X\theta})\right\|_\infty \le \eta_\theta \\ & \left\|n^{-1}\boldsymbol{X}^\top\tilde{\boldsymbol{P}}(\boldsymbol{Z}-\boldsymbol{X\theta})\right\|_\infty \le \eta'_\theta \\ & n^{-1}\boldsymbol{Z}^\top(\boldsymbol{Z}-\boldsymbol{X\theta}) \ge \bar{\eta}_\theta \\ & \|\boldsymbol{Z}-\boldsymbol{X\theta}\|_\infty \le \mu_\theta\end{array} \quad (12)$$

with tuning parameters $\eta_\theta, \eta'_\theta \asymp (\log n)\sqrt{n^{-1}\log p}$, $\mu_\theta \asymp \sqrt{\log n}$ and $0 < \bar{\eta}_\theta < \sigma_u^2$.

The first and the second constraint enable adaptive and automatic estimation of the correlation of the features of the fixed effects without prior knowledge (or correct estimation) of the correlation of the random effects. In particular, the first ensures that the gradient of the square loss is close to zero, whereas the second one looks at the reweighed gradient where the weights are defined through the matrix $\tilde{\boldsymbol{P}}$. The presence of the second constraint allows us additional flexibility with the choice of the matrix $\tilde{\boldsymbol{P}}$. We can remove the second constraint if we impose that $\lambda_{\max}(\boldsymbol{W}^\top\boldsymbol{MW})$ is bounded; however, this would restrict our choices of the matrix $\boldsymbol{M}$ and in particular it would not allow for $\boldsymbol{M} = \log(n)\mathbb{I}_n$ which we found to be beneficial whenever the

variance in the random effects is particularly large. Lastly, the last constraint excludes features $\boldsymbol{X}$ that are highly correlated with $\boldsymbol{Z}$; as our task was to remove heterogeneity (by de-correlating features), such reasoning is needed for a successful test statistic.

The Dantzig estimators can be obtained by linear programming (see, e.g., Koenker and Mizera 2014, Appendix D).

2.3. Asymptotic Inference for Linear Mixed Models

Our results are achieved under the asymptotics where $N \to \infty$ and $n_i/N \to 0$, a regime different from that used in simple linear models, and require very mild conditions on the choice of the matrix $\tilde{\boldsymbol{P}}$, that is, the matrix $\boldsymbol{M}$, which needs to have the same diagonal structure as $\boldsymbol{W}\boldsymbol{W}^\top$. However, given these high level conditions, we obtain a widely applicable result that applies to several different linear mixed models.

To define a test statistic, we focus on the moment condition (6) and in particular consider $\mathbf{A} = \tilde{\boldsymbol{P}}$. That is, the moment condition that we wish to test is now taking the form $\mathbb{E}[n^{-1}(\boldsymbol{V}-\boldsymbol{X}\boldsymbol{\gamma}^*)^\top\tilde{\boldsymbol{P}}(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\theta}^*)] = 0$ whenever the null hypothesis H_0 holds. For the alternative hypothesis, this moment takes the form $\mathbb{E}[n^{-1}(\boldsymbol{V}-\boldsymbol{X}\boldsymbol{\gamma}^*)^\top\tilde{\boldsymbol{P}}(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\theta}^*)] = \sigma_u^2 n^{-1}\text{trace}(\tilde{\boldsymbol{P}})(\beta^* - \beta_0)$.

We then proceed to form a test statistic. The moment above produces a doubly robust test statistic as long as one of two component regression models is correctly specified and assuming that there are no unmeasured confounders, giving the analyst two chances to correctly specify at least one of the regression models, (3) and (4), respectively. To standardize the test statistic appropriately, we define $\hat{\sigma}^2 = n^{-1}||\tilde{\boldsymbol{P}}(\boldsymbol{V}-\boldsymbol{X}\hat{\boldsymbol{\gamma}})||_2^2$ and $\hat{\sigma}_u^2 = n^{-1}||(\boldsymbol{Z}-\boldsymbol{X}\hat{\boldsymbol{\theta}})||_2^2$. Whenever the models are correctly specified, they produce residual variance estimates. However, the quality of testing does not rely on them being consistent estimators; in fact, as the proxy log-likelihood is not exact, this will not be achievable; we show in Lemma B.7, in the supplementary materials, that $\hat{\sigma}^2 = \sigma_\epsilon^2 n^{-1}\text{trace}(\tilde{\boldsymbol{P}}\boldsymbol{P}^{-1}\tilde{\boldsymbol{P}}) + o_P(1)$.

Now, we proceed to define the test statistic as

$$T_n = \frac{n^{-1/2}(\boldsymbol{Z}-\boldsymbol{X}\hat{\boldsymbol{\theta}})^\top\tilde{\boldsymbol{P}}(\boldsymbol{V}-\boldsymbol{X}\hat{\boldsymbol{\gamma}})}{\hat{\sigma}_u\hat{\sigma}}. \tag{13}$$

The doubly robust test statistic combines the initial model under the null hypothesis with a regression model of the relationship between covariates and the feature related to each of the parameters of interest in such a way that, as long as either the initial model or the feature regression model is correctly specified, the effect of the initial null hypothesis on the transformed moment is correctly estimated. In the case of linear mixed models, the initial model is only correctly specified when the variance parameters $\boldsymbol{\Psi}$ are known, making the test statistic above particular useful for cases when there is no such knowledge.

Our first result is that the test statistic has asymptotically a standard normal distribution whenever the null hypothesis holds (see Theorem 3). For this result we do not require sparsity of the fixed effects, but rather sparsity of the precision matrix of the design matrix of the fixed effects. However, we do not require any consistent estimation of the covariance parameters Ψ. To obtain the asymptotic distribution under the alternative hypothesis, $H_1 : \beta^* = \beta_0 + n^{-1/2}h$, we need to assume that the vector of nuisance parameters $\boldsymbol{\gamma}^*$ is sparse (or approximately) with a small number of nonzero elements, although our simulations show that the power is preserved even if the fixed effects have as many as n nonzero elements (see Theorem 4). To our knowledge, all existing results regarding inference of linear mixed models require consistent estimation of the covariance parameters and low-dimensionality of the fixed effects vector. Although some empirical work has demonstrated that inconsistent variance estimation does not effect Type I error control (in low-dimensional settings), theoretical guarantees were never established.

3. Asymptotic Theory for Linear Mixed Models

To use the test statistics to provide formally valid statistical inference, we need an asymptotic normality theory in the desired asymptotics. We begin by precisely describing the asymptotics under the null hypothesis. We then proceed to describe the asymptotics under the alternative hypothesis

$$H_1 : \beta^* = \beta_0 + hn^{-1/2}. \tag{14}$$

Before stating our theoretical results, we give some definitions.

Definition 1 (P-condition). We say that a symmetric semidefinite positive matrix $\boldsymbol{A}$ satisfies the *P-condition* if it has the same block diagonal structure as $\boldsymbol{W}^\top\boldsymbol{W}$ and if each element of $\boldsymbol{A}$ is $\mathcal{O}(\log(n))$.

In Lemma B.4, we show that $\tilde{\boldsymbol{P}} = (\boldsymbol{I}_n + \boldsymbol{W}\boldsymbol{M}\boldsymbol{W}^\top)^{-1}$ satisfies the P-condition for any matrix $\boldsymbol{M}$ having the same block diagonal structure as $\boldsymbol{W}^\top\boldsymbol{W}$. Furthermore, it holds that $\boldsymbol{P}$ and $\boldsymbol{P}^{-1}$ satisfy the P-condition.

Next, we require that the errors $\boldsymbol{\epsilon}$ in our model have sub-Gaussian tails. A test statistic, as defined in Section 2.3, can be used to give honest p-values or confidence intervals; in Section 4 we illustrate its good properties even when the error is not sub-Gaussian but exhibits heavy-tails.

Definition 2. A random variable X is said to have an exponential-type tail with parameters (b, γ) if $\forall x > 0$, $\mathbb{P}(|X| > x) \leq \exp[1 - (x/b)^\gamma)]$. Furthermore, a random variable X is sub-Gaussian if it has an exponential-type tail with parameters $(b, 2)$.

To guarantee consistency, we also need to enforce that the design matrix of the fixed effects satisfies some regularity conditions. The conditions are standard in the high-dimensional literature and ensure that the population feature covariance matrix is a well-conditioned matrix. Here, we follow Rudelson and Zhou (2013), and achieve this effect by enforcing sub-Gaussianity in the design in the following condition. Let $C_{\min}$, $C_{\max}$, and κ denote positive constants.

Condition 1. The matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}^\top\boldsymbol{X}]/n \in \mathbb{R}^{(p-1)\times(p-1)}$ is such that $C_{\min} \leq \sigma_{\min}(\boldsymbol{\Sigma}) \leq \sigma_{\max}(\boldsymbol{\Sigma}) \leq C_{\max}$ with $\sigma_{\min}(\boldsymbol{\Sigma})$ and $\sigma_{\max}(\boldsymbol{\Sigma})$ the minimal and maximal singular values of $\boldsymbol{\Sigma}$. The vectors $\boldsymbol{\Sigma}^{-1/2}x_i$, $\boldsymbol{\epsilon}$, and $\boldsymbol{u}$ are centered with sub-Gaussian norms upper bounded by κ. Moreover, $\sigma_u, \sigma_\epsilon \in [C_{\min}, C_{\max}]$.

The elements of the symmetric and invertible matrix $\boldsymbol{\Psi}$ and of the deterministic design matrix $\boldsymbol{W}$ are bounded.

The remaining definition is more technical. We use a regularity condition to control the shape of the correlation in the features and in the variance covariance matrix $\boldsymbol{\Psi}$. Such condition is used regularly in high-dimensional estimation and inference (see Bickel, Ritov, and Tsybakov 2009).

Definition 3. We say that the restricted eigenvalue condition holds for a triplet $(s, \kappa, \boldsymbol{X})$ if

$$\min_{\substack{J_0 \subseteq \{1,\dots,p-1\} \\ |J_0| \leq s}} \min_{\substack{\delta \neq 0 \\ \left\|\delta_{J_0^c}\right\|_1 \leq \left\|\delta_{J_0}\right\|_1}} \frac{\|\boldsymbol{X}\delta\|_2}{\sqrt{n}\left\|\delta_{J_0}\right\|_2} \geq \kappa.$$

Rudelson and Zhou (2013) showed that the restricted eigenvalue condition holds with large probability if the sample size n is large enough and the rows of the design have sub-Gaussian distributions.

We note that our test statistic (13) does not depend on the inverse of $\boldsymbol{\Psi}$. So although we started this section with the nonsingularity assumption of $\boldsymbol{\Psi}$, in practice our method can be directly applied even when noise random effects exist.

Given these preliminaries, we state our main result on the asymptotic normality of the test statistic T_n. As discussed previously, we require that the sample size scales as $\sqrt{n}/\log(p) = o(1)$ as $N \to \infty$. If the subsample size grows slower than this, the test statistic will still be asymptotically normal, but the it may be asymptotically biased. Although we treat n, s and p as functions of N, for clarity, we state the following result without this explicit dependence.

Theorem 1. Suppose that we have N independent observations $(\mathbf{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \mathbf{W}_i) \in \mathbb{R}^{n_i} \times \mathbb{R}^{n_i \times (p-1)} \times \mathbb{R}^{n_i} \times \mathbb{R}^{n_1 \times q}$. Suppose moreover that the features satisfy Condition 1. Let the matrix $\tilde{\boldsymbol{P}}$ be such that Definition 1 holds. Given this data-generating process, let $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\theta}}$ be estimators of the nuisance parameters and the feature dependence as defined in (10) and (12). Finally, suppose that the sample size scales as $n^{1/2}/\log(p)/\log(n) = o(1)$ and $\max_i n_i/N = o(1)$ and that the models are normalized so that $||\boldsymbol{\gamma}^*||_2 = O(1)$ and $||\boldsymbol{\theta}^*||_2 = O(1)$. Then,

$$T_n \overset{d}{=} \mathcal{N}(0,1) + \sqrt{n}(\beta^* - \beta_0)\frac{n^{-1}\mathrm{trace}(\tilde{\boldsymbol{P}})}{\sqrt{n^{-1}\mathrm{trace}(\tilde{\boldsymbol{P}}\boldsymbol{P}^{-1}\tilde{\boldsymbol{P}})}} + o_P(r_n),$$

where $r_n = n^{1/2}\eta_\theta\eta_\gamma \left\|\boldsymbol{\theta}^*\right\|_0 \kappa^{-2}\bar{\eta}_\gamma^{-1}\sigma_\epsilon/\sigma_u$ whenever $\beta^* = \beta_0$ and $r_n = n^{1/2}\eta_\gamma \left\|\boldsymbol{\gamma}^*\right\|_0 \kappa^{-2}\eta'_\theta\sigma_\epsilon/\sigma_u + \eta_\theta\sqrt{\left\|\boldsymbol{\theta}^*\right\|_0}\sigma_\epsilon/\sigma_u + \sqrt{n}(\beta^* - \beta_0)N/n\sigma_\epsilon/\sigma_u$ whenever $\beta^* \neq \beta_0$.

The construction of p-values is based on the asymptotic distribution of the test statistic T_n. For the null hypothesis $H_0 : \beta^* = \beta_0$, we define the p-value for the two-sided alternative as $P_0 = 2(1 - \Phi(|T_n(\beta_0)|))$. Of course, we could also consider one-sided alternatives with an obvious modification. Whenever the conditions of the Theorem 1 hold, then for any $0 < \alpha < 1$,

$$\lim_{n\to\infty} \sup P\left[P_0 \leq \alpha\right] = \alpha \text{ if } H_0 \text{ holds}.$$

Furthermore, for any sequence $a_n \to 0$ ($n \to \infty$) which converges sufficiently slowly, the statements also hold when replacing α by a_n. A discussion about detection power of the method is given in Section 3.

The proof of Theorem 1 is organized as follows. In Section 3.2, we provide bounds for the l_1 norm error in estimation of both $\boldsymbol{\gamma}^*$ and $\boldsymbol{\theta}^*$, while Section 3.3.1 studies the sampling distributions of the test statistic under the null hypothesis and establishes asymptotic Gaussianity. Given a subsampling rate and sparsity requirements in the two models, (3) and (4), we showcase the asymptotic distribution under the alternative hypothesis in Section 3.3.2. Before beginning the proof, however, we relate our result to existing results about linear mixed models in Section 3.1.

3.1. Theoretical Background

There has been considerable work in understanding the theoretical properties of linear mixed models. The convergence and consistency properties of estimation have been studied by, among others, McCulloch (1997), McGilchrist (1994), and Lindstrom and Bates (1988) in low-dimensional setting, Goeman, Van Houwelingen, and Finos (2011) with growing dimensions, and Schelldorfer, Buhlmann, and van de Geer (2011) in high-dimensional setting. Meanwhile, the inference has been analyzed by Breslow and Clayton (1993) in low-dimensional setting. However, to our knowledge, our Theorem 1 is the first result establishing conditions under which inference in linear mixed models can be performed despite the presence of a large number of fixed effects and taking into account model selection. Moreover, Theorem 1 establishes this result without even resorting to a proper (let alone consistent) estimation of the random effects therefore enabling inference in linear models with misspecified random effects (the structure or distribution); we believe that such result is unique in inferential theory of linear mixed models.

Probably the closest existing result is that of Fan and Li (2012), who showed that selection of fixed effects can be achieved without resorting to proper estimation of random effects. They establish model selection consistency results under conditions similar to ours; however, we note that they impose a slightly more restrictive choice of the matrix $\tilde{\boldsymbol{P}}$ whereas we allow more general choices—for example, choice of $\boldsymbol{M}$ as an identity matrix is not allowed in their work, but it is in ours. Moreover, the authors therein do not discuss asymptotic distributions and hypothesis testing problems. Thus, their results cannot be used for valid asymptotic statistical inference about fixed effects.

Inference for linear mixed models that is robust to the normality assumption in random effects or the model error has been studied exclusively in low-dimensional setting in Zhang and Davidian (2001), who showed that normality can be substituted with a broader class of distributions and an EM algorithm can be used for estimation. However, their results still heavily depend on a correct likelihood specification. Identical conclusions hold for a class of semiparametric linear mixed models introduced in Lachos, Ghosh, and Arellano-Valle (2010) or a linear mixed model with random mixture of normals (Verbeke and Lesaffre 1996; Song, Zhang, and Qu 2007). See Heagerty and Kurland

(2001) for an illustration of the effects of inconsistent variance estimation for the likelihood based inference on fixed effects. Additional work on this topic (Litiére, Alonso, and Molenberghs 2007; Hobert and Casella 1996) had only highlighted connections to statistical tests and the effects of misspecification of the random effects, but did not establish any formal justification for it.

3.2. Estimation Properties

We start by bounding the bias of the newly introduced high-dimensional estimators. Our approach relies on showing that the true parameter vector lies in the constraint set of the problems (10) and (12) for appropriate choices of the tuning parameters defined therein. A sparsity assumption then allows us to bound the bias. We show that the error of estimation is optimal and is not effected by the misspecification of the random effects. To state the result, define $||\boldsymbol{\gamma}^*||_0$ as the number of nonzero elements of a vector $\boldsymbol{\gamma}^*$.

Theorem 2. Suppose $\tilde{\boldsymbol{P}} \in \mathbb{R}^{n\times n}$ satisfies the P-condition. Suppose that Condition 1 holds and that there exists $\kappa > 0$ such that the restricted eigenvalue condition holds for $(||\boldsymbol{\gamma}^*||_0, \kappa, \tilde{\boldsymbol{P}}^{1/2}\boldsymbol{X})$. Then there exist constants C_1, C_2, $c_0 > 0$ such that for any $\eta_\gamma \geq C_1 \log(n)\sqrt{n^{-1}\log p}$, $\mu_\gamma \geq C_2\sqrt{\log n}$ and $\bar{\eta}_\gamma, \in (0, n^{-1}\text{trace}(\sigma_\epsilon^2 \tilde{\boldsymbol{P}}\boldsymbol{P}^{-1}) - c_0)$, for a large enough constant C_3 it holds with probability $1 - p^{-C_3}$

$$\left\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\right\|_1 \leq 8\eta_\gamma \left\|\boldsymbol{\gamma}^*\right\|_0 \kappa^{-2}, \text{ and} \tag{15}$$

$$n^{-1}\left\|\tilde{\boldsymbol{P}}^{1/2}\boldsymbol{X}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\right\|_2^2 \leq 16\eta_\gamma^2 \left\|\boldsymbol{\gamma}^*\right\|_0 \kappa^{-2}. \tag{16}$$

This theorem then directly translates into a bound on the bias in estimation. Namely, with $\eta_\gamma \asymp \log(n)\sqrt{n^{-1}\log p}$, $\mu_\gamma \asymp \sqrt{\log n}$ and $0 < \bar{\eta}_\gamma < n^{-1}\text{trace}(\sigma_\epsilon^2 \tilde{\boldsymbol{P}}\boldsymbol{P}^{-1})$ the estimator $\hat{\boldsymbol{\gamma}}$ of a s sparse vector $\boldsymbol{\gamma}^*$,

$$\left\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\right\|_1 = \mathcal{O}_P(s\log(n)\sqrt{\log p/n}).$$

The conditions required for the above result seem very mild. Finite sample oracle risk properties have been established in Schelldorfer, Buhlmann, and van de Geer (2011); however, the result therein crucially depends on correct estimation of variance components. In contrast, our results hold for a wide variety of choices of variance estimators; in particular, it remains valid for even non-consistent estimators—the only assumption is on the structure of the matrix **P** illustrated through the P-condition. Furthermore, the rate above matches the rate obtained by the mle estimator of Schelldorfer, Buhlmann, and van de Geer (2011) highlighting excellent robustness properties of the proposed estimator $\hat{\boldsymbol{\gamma}}$ and indicating optimality at estimation.

3.3. Testing Properties

Our analysis proceeds in two steps: the first discusses asymptotic normality of the test statistic whenever the null hypothesis holds, whereas the second discusses the case under the alternative hypothesis.

3.3.1. Size

Analyzing specific linear mixed models can be challenging especially if the model (1) is not fully or correctly specified. Below we establish asymptotic normality of the proposed test statistic. The result is independent of the size of the sparsity of the nuisance parameters as long as the signal-to-noise ratio is not unbounded or equivalently as long as l_2 norm of the nuisance parameters is bounded. A collection of very many weak signals or a mixture of a large number of weak and a small number of strong signals would satisfy this setting. Moreover, it does not require consistent estimation of the variance components.

Theorem 3. Suppose $\tilde{\boldsymbol{P}} \in \mathbb{R}^{n\times n}$ satisfies the P-condition. Suppose that Condition 1 holds and that there exists $\kappa > 0$ such that the restricted eigenvalue condition holds for $(||\boldsymbol{\theta}^*||_0, \kappa, \tilde{\boldsymbol{P}}^{1/2}\boldsymbol{X})$. Furthermore, assume that $\left\|\boldsymbol{\theta}^*\right\|_0 = o(\sqrt{n}/\log(p)/\log(n))$ and $\|\boldsymbol{\gamma}^*\|_2 = \mathcal{O}(1)$, $\left\|\boldsymbol{\theta}^*\right\|_2 = \mathcal{O}(1)$. Then, under the null hypothesis $H_0 : \beta_* = \beta_0$, it holds that $T_n \xrightarrow{d} \mathcal{N}(0,1)$.

Theorem 3 establishes the asymptotic distribution of our test statistics under the null hypothesis; it does not require minimal signal strength in the model (1) or an irrepresentable condition. This finding is further illustrated numerically in Section 4 where we observe stable control over the Type I error rate for a wide range of sizes $||\boldsymbol{\gamma}^*||_0$.

Furthermore, using some arguments of the proof of Theorem 4, interchangeability of sparsity conditions can be shown: if $\boldsymbol{\gamma}^*$ is sparse then $\boldsymbol{\theta}^*$ does not need to be sparse. Note that the $o(\sqrt{n}/\log(p)/\log(n))$ sparsity rate is up to a $\log(n)$ term matching those of simple linear models (see, e.g., Van de Geer et al. 2014; Javanmard and Montanari 2014b). The additional $\log(n)$ term is needed in controlling the random effects and can be though of as a price to pay for being able to provide optimal estimation despite the possibly incorrect specification of the random effects.

Besides providing the size property of the T_n statistic, Theorem 3 can also be used to construct confidence intervals. The $1-\alpha$ confidence interval for β^* can be defined as

$$\left\{\beta : 1 - \Phi^{-1}(1-\alpha/2) \leq T_n(\beta) \leq \Phi^{-1}(1-\alpha/2)\right\},$$

where Φ is the standard Gaussian cumulative distribution function and where $T_n(\beta)$ is the test statistic derived under the null hypothesis $\beta^* = \beta$.

3.3.2. Power

In the previous section, we showed that the Type I error is asymptotically equal to α for a given level $\alpha \in (0,1)$. In this section, we show that the test statistic furthermore has tight control of the Type II error and preserves power asymptotically while allowing inconsistent variance estimation. In this regard, we believe that our result stands out in the existing literature. In addition, we quantify the asymptotic efficiency loss due to misspecification of the random effects.

Theorem 4. Suppose $\tilde{\boldsymbol{P}} \in \mathbb{R}^{n\times n}$ satisfies the P-condition, Condition 1 holds and that there exists $\kappa > 0$ such that the restricted eigenvalue condition holds for $(||\boldsymbol{\theta}^*||_0 \vee ||\boldsymbol{\gamma}^*||_0, \kappa, \tilde{\boldsymbol{P}}^{1/2}\boldsymbol{X})$. Furthermore, assume that both $\left\|\boldsymbol{\theta}^*\right\|_0 = o(\sqrt{n}/\log(p)/\log(n))$

and $\|\boldsymbol{\gamma}^*\|_0 = o(\sqrt{n}/\log(p)/\log(n))$ and that the normalization $\|\boldsymbol{\gamma}^*\|_2 = \mathcal{O}(1)$, $\|\boldsymbol{\theta}^*\|_2 = \mathcal{O}(1)$ holds. Then, under the alternative hypothesis H_1 in (14), it holds that with $n \to \infty$

$$T_n \stackrel{d}{=} \mathcal{N}(0,1) + h\sigma_u\sigma_\epsilon^{-1}\frac{n^{-1}\text{trace}(\tilde{\boldsymbol{P}})}{\sqrt{n^{-1}\text{trace}(\tilde{\boldsymbol{P}}\boldsymbol{P}^{-1}\tilde{\boldsymbol{P}})}} + o_P(1).$$

Theorem 4 establishes the power properties of our test and requires both $\boldsymbol{\gamma}^*$ and $\boldsymbol{\theta}^*$ to be sparse with a $o(\sqrt{n}/\log(p)/\log(n))$ sparsity rate. The deviation term depends on the choice of the matrix $\tilde{\boldsymbol{P}}$ and by the Cauchy–Schwarz inequality it can be shown that

$$\frac{n^{-1}\text{trace}(\tilde{\boldsymbol{P}})}{\sqrt{n^{-1}\text{trace}(\tilde{\boldsymbol{P}}\boldsymbol{P}^{-1}\tilde{\boldsymbol{P}})}} \leq \sqrt{n^{-1}\text{trace}(\boldsymbol{P})}.$$

This implies that using the proxy matrix $\tilde{\boldsymbol{P}}$ instead of the true unknown matrix $\boldsymbol{P}$ leads to an asymptotic loss of power.

If we consider the particular, naive choice of the proxy matrix $\boldsymbol{M} = \boldsymbol{0}_{Nq\times Nq}$ then $\tilde{\boldsymbol{P}}$ reduces to the identity matrix $\boldsymbol{I}_{Nq}$. In that case the method reduces to the method entirely based on simple linear model and leads to a significant loss of power. For a simple choice of the matrix $\boldsymbol{M} = c\boldsymbol{I}_{Nq}$ where c is a constant positive number, the matrix $\tilde{\boldsymbol{P}}$ takes the form $(\boldsymbol{I}_n + c\boldsymbol{W}\boldsymbol{W}^\top)^{-1}$.

Using Theorem 4 we can explicitly derive the expression for power of the T_n statistics under nominal level α as

$$\text{Power} = 2 - \Phi\left(\Phi^{-1}(1-\frac{\alpha}{2}) - D_n(h)\right) + \Phi\left(\Phi^{-1}(1-\frac{\alpha}{2}) + D_n(h)\right),$$

where $D_n(h) = h\sigma_u\sigma_\epsilon^{-1}\frac{n^{-1}\text{trace}(\tilde{\boldsymbol{P}})}{\sqrt{n^{-1}\text{trace}(\tilde{\boldsymbol{P}}\boldsymbol{P}^{-1}\tilde{\boldsymbol{P}})}}$.

Furthermore, the length of the $1-\alpha$ confidence interval of β^* constructed from Theorem 3 is given by $2n^{-1/2}|h_\alpha|$ for h_α satisfying $2\Phi\left(\Phi^{-1}(1-\alpha/2) - D_n(h_\alpha)\right) = 1-\alpha$ which can easily be estimated numerically.

We also note that the result can be extended to include fully dense models where $||\boldsymbol{\gamma}^*||_\infty$ is not allowed to grow rapidly with n but the l_0 norm can be as large as p; or mixture settings with many small elements and only a small number of strong elements (here the sum of the small elements would be allowed to grow with p). We conjecture that the power function will look the same as it does in Theorem 4.

4. Numerical Experiments

In observational studies, accurate construction of p-values requires to overcome three potential sources of bias. First, there is the initial linear mixed model and in particular the size of the sparsity of the nuisance parameters $\boldsymbol{\gamma}^*$. We need to make sure that our test is reasonably stable whatever the structure of $\boldsymbol{\gamma}^*$ is. Then, second, there is the underlying precision matrix of the design matrix. Here a bias can be introduced with presence of large feature correlations and/or dependencies. Third, we need to make sure that the test is reasonably stable regardless of the exact structure of the random effects; we need to make sure that the test is not biased by changing matrix $\boldsymbol{\Psi}$, that is, $\boldsymbol{\psi}$. The simulations here aim to test the ability of the introduced test to respond to all three of these factors.

4.1. Models and Designs

We consider the mixed model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{W}\boldsymbol{b} + \epsilon$$

with $n = 200$ observations coming from either $N = 50$ different groups of size $n_i = 4$ or $N = 20$ groups of size $n_i = 10$. The vector $\boldsymbol{\beta}^*$ of fixed effects is of length $p = 500$. The components of $\boldsymbol{\epsilon}$ are independent and come from a standard Gaussian distribution. The random effect vector is defined as $\boldsymbol{b} = (\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_N^\top)^\top \in \mathbb{R}^{qN}$ with $\boldsymbol{b}_i \sim \mathcal{N}_q(0, \boldsymbol{\psi})$ for $i = 1, \ldots, N$. We define $\boldsymbol{W} = \text{diag}(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_N) \in \mathbb{R}^{n\times qN}$ with the stack matrix $(\boldsymbol{W}_1^\top, \ldots, \boldsymbol{W}_N^\top)^\top \in \mathbb{R}^{n\times q}$ consisting of the first q columns of $\boldsymbol{X}$. We consider three models:

- *Model 1*: The rows of $\boldsymbol{X}$ come from a multivariate normal distribution with zero mean and Toeplitz covariance matrix $\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma}_{ij} = (-0.5)^{|i-j|}$. The components of $\boldsymbol{\beta}^*$ are generated as follows: $\boldsymbol{\beta}^* = 5\boldsymbol{a}/\|\boldsymbol{a}\|_2$ with $\boldsymbol{a} = (a_1, \ldots, a_p)$, where a_j is generated from a uniform distribution on $(0, 1)$ if $j \leq 3s/2$ and $j/3$ is not an integer; otherwise, $a_j = 0$. We then have $\|\boldsymbol{\beta}\|_0 = s$. The model contains $q = 2$ random effects with covariance matrix $\boldsymbol{\psi} = \text{diag}(0.56, 0.56)$.
- *Model 2*: The setting is the same as in Model 1 except that $q = 3$ and $\boldsymbol{\psi} = \text{diag}(3, 3, 2)$. For Model 2(b) in addition to the setting of Model 2, $\text{cov}(b_{i1}, b_{i2}) = 0.3$, $\text{cov}(b_{i1}, b_{i3}) = 0.1$ and $\text{cov}(b_{i2}, b_{i3}) = 0.2$.
- *Model 3*: The setting is the same as in Model 1 except for $\boldsymbol{\Sigma}$ which is such that $\boldsymbol{\Sigma}_{ij} = 1$ if $i = j$, $\boldsymbol{\Sigma}_{ij} = -\rho/(1+\rho^2)$ if $|i-j| = 1$, and $\boldsymbol{\Sigma}_{ij} = 0$ otherwise, where ρ varies.

As noted in the Appendix, the Toeplitz design implies that the underlying correlation model (4) is sparse with two nonzero components being equal (for $\rho = -0.5$) to –0.4. Thus, for Models 1 and 2, the vector $\boldsymbol{\theta}^*$ is sparse. Conversely, Model 3 is such that that the rows of Σ are sparse, while the row of its inverse are not sparse. This in turn implies that $\boldsymbol{\theta}^*$ in (4) is a dense vector. Regarding the random effect part, similar choices of q and $\boldsymbol{\psi}$ are made by Schelldorfer, Buhlmann, and van de Geer (2011).

4.2. Tuning Parameters

Tuning parameters need to be chosen in optimization problems (10) and (12). Based on our investigation, we suggest the following choices. The sensitivity of our procedure to the choice of the tuning parameters is given in Table 5. Similarly to Zhu and Bradic (2018), we choose $\bar{\eta}_\gamma = 0.05\ \boldsymbol{V}^\top\tilde{\boldsymbol{P}}\boldsymbol{V}/n$. We define $\eta_\gamma = \sqrt{0.5n^{-1}\log p}\ \hat{\sigma}$ and $\mu_\gamma = 4\sqrt{\log n}\ \hat{\sigma}$ with $\hat{\sigma} = \left\|\tilde{\boldsymbol{P}}(\boldsymbol{V} - \boldsymbol{X}\boldsymbol{\gamma}_{\text{init}})\right\|_2$ where $\boldsymbol{\gamma}_{\text{init}}$ is obtained by a linear estimation (without the random effects) with the scaled lasso. Such an initial estimator of $\boldsymbol{\gamma}$ is also used by Rohart, San-Cristobal, and Laurent (2014). The choice of the tuning parameters $\bar{\eta}_\theta$, η_θ, μ_η, and η'_θ is done in a similar way.

Regarding the choice of the matrix $\tilde{\boldsymbol{P}} = (\boldsymbol{I}_n + \boldsymbol{W}\boldsymbol{M}\boldsymbol{W}^\top)^{-1}$ serving as a proxy of $\boldsymbol{P} = (\boldsymbol{I}_n + \sigma_\epsilon^{-2}\boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}^\top)^{-1}$, we show in Lemma B.4 (see the Supplementary Materials) that there is a large pool of possible choices guaranteeing good

asymptotic results. Fan and Li (2012) suggest to use $\boldsymbol{M} = (\log n)\boldsymbol{I}_{Nq}$. In the present simulation study, we use the slightly more elaborate $\boldsymbol{M} = \sigma^{-2}_{\epsilon,\text{init}}\text{diag}(\boldsymbol{\psi}_{\text{init}},\ldots,\boldsymbol{\psi}_{\text{init}})$ where initial estimators are derived similarly as in Rohart, San-Cristobal, and Laurent (2014): using $\boldsymbol{\gamma}_{\text{init}}$ obtained by a linear model estimation with the scaled lasso, we compute $\sigma^{2[-1]}_{\epsilon} = \frac{1}{n-\text{df}}\|\boldsymbol{V} - \boldsymbol{X}\boldsymbol{\gamma}_{\text{init}}\|^2_2$, $\psi_{\text{init}} = \frac{0.4}{q}\sigma^{2[-1]}_{\epsilon}\boldsymbol{I}_q$, and $\sigma^2_{\epsilon,\text{init}} = 0.6\sigma^{2[-1]}_{\epsilon}$.

4.3. Numerical Results

We test $H_{0,j} : \beta^*_j = \beta_{0,j}$ for $j = 3, 4, 100$ and study the behavior of our procedure. These values of j correspond to different types of coefficients; components 3 and 100 are not active while component 4 is active. Furthermore, in the Toeplitz case, component 100 is almost not correlated with any of the active variables. We set $H_{1,j} : \beta^*_j = \beta_{0,j} + hn^{-1/2}$ where the situation $h = 0$ corresponds to size properties while $h \neq 0$ corresponds to power properties. All our results are based on 1000 replications and on a 5% nominal level. To the best of our knowledge, our procedure is the first one to do hypothesis testing for high-dimensional mixed models which makes it hard to find competitors. We compare our mixed model method to the linear model method of Zhu and Bradic (2018), that is, to a model that ignores the random effects.

4.3.1. Gaussian Mixed Model

In Figure 1, we report size and power properties of our procedure for Model 1 and for different components of $\boldsymbol{\beta}^*$. In Figure 1(a), we observe that the false discovery rate is close to the 5% nominal level whatever the sparsity level. Our procedure can handle dense $\boldsymbol{\beta}^*$ and this is one of its major features. Sparsity is required only for either $\boldsymbol{\theta}^*$ or $\boldsymbol{\beta}^*$ to obtain reliable results under the null hypothesis. Note that under the alternative hypothesis $h \neq 0$, our theoretical results require both of them to be sparse, see Theorem 4. In Figure 1(b), we provide the power of our test for Model 1 in a sparse model ($s = 5$). Interestingly, the probability to reject the null hypothesis is not symmetric in h. This can be explained by the fact that we use a penalized estimator which shrinks coefficients toward zero. The correlation between the components compensates (here for h negative) or amplifies (here for h positive) the bias of the estimator, which leads to an asymmetry of the results.

In Figure 2, we give size and power properties of our procedure for Model 3 in which $\boldsymbol{\theta}^*$ is dense, that is, $s_\theta = p - 1$ in this particular case. We vary the size of the correlation ρ with larger values corresponding to larger size of the coefficients of $\boldsymbol{\theta}^*$. We observe that the proposed test is remarkably resilient and controls Type I error in finite samples. We observe that when $\boldsymbol{\gamma}^*$ is not sparse, which here corresponds to sparsity larger than 5 (note that $\sqrt{n}/\log(p) \approx 5$ for $n = 200$, $p = 500$), our method is not guaranteed to control Type I error; however, we see that our method strikingly controls errors even in such cases whenever $\rho < 0.5$. It is worth pointing that no method is expected to perform well when all elements are dense and correlation is high. Hence, we take this behavior to be close to optimal.

In Table 2, we compare our method, for Models 1 and 2, to the method of Zhu and Bradic (2018) designed for linear models and which thus does not take into account the group structure of the mixed model. For both models, the probability to reject the null hypothesis is closer to the 5% nominal level when the null hypothesis holds ($h = 0$) and is larger under the alternative hypothesis ($h \neq 0$), compared to the linear model procedure.

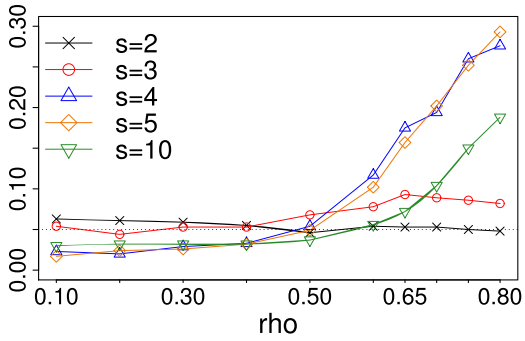

Figure 2. Empirical rate of rejection of the true null hypothesis at a 5% nominal level for Model 3 with dense vector $\boldsymbol{\theta}^*$. We vary the level of sparsity of $\boldsymbol{\gamma}^*$ and correlation level ρ. Hypothesis testing is performed on the second component of $\boldsymbol{\beta}^*$. Empirical Type I errors are plotted for different sparsity levels: black crosses 2, red circles 3, blue triangles 4, orange diamonds 5, green upside-down triangles 10.

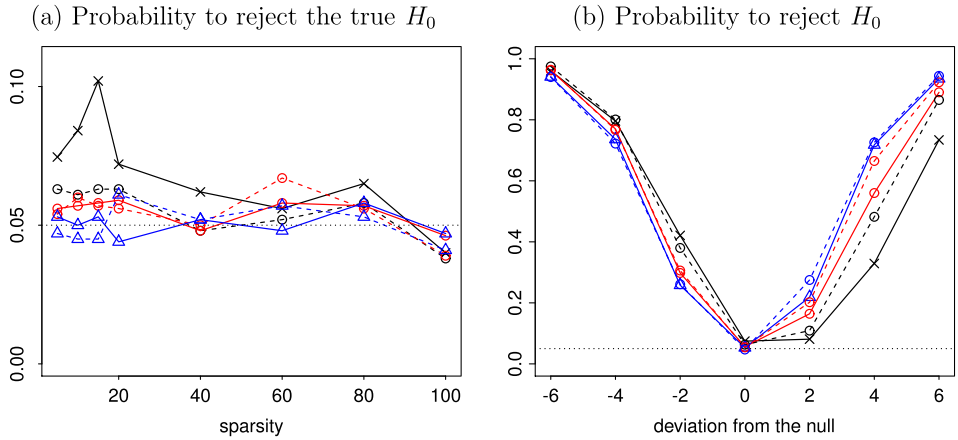

Figure 1. Empirical rate of rejection of the null hypothesis under (a) the null hypothesis, and (b) alternative hypothesis, at a 5% nominal level for Model 1. As described in the text, hypothesis tests are performed on different components of β^*: the third component (black crosses), the fourth component (red circles), and the 100th component (blue triangles). Full lines are for $n_i = 4$, dashed lines for $n_i = 10$.

Table 2. Probability of rejecting the null hypothesis as a function of the deviation from the null.

	Probability to reject the null hypothesis at a 5% level					
	Model 1			Model 2		2(b)
	LM	MM	MM(a)	LM	MM	MM
$h = -6$	0.85	0.96	0.96	0.46	0.65	0.60
$h = -4$	0.60	0.77	0.77	0.30	0.39	0.36
$h = -2$	0.29	0.30	0.31	0.19	0.16	0.14
$h = 0$	0.11	0.06	0.05	0.12	0.05	0.04
$h = 2$	0.11	0.16	0.20	0.07	0.05	0.06
$h = 4$	0.28	0.56	0.66	0.07	0.18	0.23
$h = 6$	0.55	0.89	0.92	0.13	0.42	0.47

NOTE: Nominal level is taken to be 5%. *LM* ignores the random effects and applies the linear model procedure of Zhu and Bradic (2018) while *MM* consists in applying our mixed model procedure. The 4th component of $\boldsymbol{\beta}^*$ is tested. Settings are $n = 200, p = 500, s = 5, n_i = 4$ except for (a) $n_i = 10$. Model 1 contains 2 random effects and Model 2 contains 3 random effects with large magnitude. For (b) the random effects are correlated.

4.3.2. Model With Heavy-Tailed Design

In the next example, we consider a model that departs from normality assumptions. Parameters choices are done in the same way as in Model 1: $n = 200$, $N = 50$, $p = 500$, and $q = 2$.

- *Model 4*: The setting is the same as in Model 1 except that the entries of $\boldsymbol{\Sigma}^{-1/2}x_i$ and of $\boldsymbol{\epsilon}$ are generated from a Student's t-distribution with 10 degrees of freedom.
- *Model 5*: The setting is the same as in Model 1 except that the entries of $\boldsymbol{\Sigma}^{-1/2}x_i$ and of $\boldsymbol{\epsilon}$ are generated from a Student's t-distribution with 3 degrees of freedom.

We perform hypothesis tests for the fourth component of $\boldsymbol{\beta}^*$ for Models 1, 4, and 5 and report empirical sizes and powers in Figure 3. We observe empirical coverages close to the 5% nominal level for the three models under the null hypothesis whatever the sparsity level is. We also observe that the presence of heavier tailed errors results in a decrease of power.

4.3.3. Effects of Misspecifying the Random Effects Structure

While the previous simulation settings already included the case of ignoring the random effects structure, we here investigate the effects of misspecifying the random effects structure in a hierarchical mixed model with two random effects. We generate from a hypothetical clinical trial where there are 5 doctors each treating 10 patients with 3 repeated measurements per patient. This corresponds to a hierarchical mixed model with a random effect for doctor, assumed in the simulation $N(0, 0.5)$, and a random effect for the patient, assumed $N(0, 1.2)$. We took the error $N(0, 1)$ and kept the covariate design as in Model 1 with $p = 500$ and $s = 5$. To investigate the misspecification of the random effect structure we fit the model in three ways: (i) assuming the correct 2-level model, (ii) only fitting a random effect for each of the 50 patients but ignoring that patients are clustered by doctor, and (iii) only fitting a random effect per doctor and ignoring the effect that there might be correlation due to the repeated measurements per patient. The latter is a more severe misspecification. The simulation results are summarized in Table 3 and Figure 4. We observe that the misspecification has no effect on the level of the test, while the more severe misspecification leads to some loss in power. Failing to include the random effect per doctor is negligible in this setting.

4.3.4. Choice of the Proxy Matrix

The choice of the proxy matrix $\tilde{\boldsymbol{P}} = (\boldsymbol{I}_n + \boldsymbol{W}\boldsymbol{M}\boldsymbol{W}^\top)^{-1}$ has an influence on the power of the test. Our default choice, presented in Section 4.2 and inspired by Rohart, San-Cristobal, and Laurent (2014) consists in working with $\boldsymbol{M} = \sigma^{-2}_{\epsilon,\text{init}}\text{diag}(\boldsymbol{\psi}_{\text{init}}, \ldots, \boldsymbol{\psi}_{\text{init}})$. With these initial estimators, this is equivalent to work with $M = \frac{2}{3q}\boldsymbol{I}_{Nq}$. A second choice, proposed by Fan and Li (2012), is to work with $\boldsymbol{M} = \log n \boldsymbol{I}_{Nq}$. Simulation results for those choices of the proxy matrix are reported in Table 4 for Models 1 and 2. We first observe that the two choices provide the 5% nominal level under the null hypothesis. Secondly, we observe slight differences in terms of power which can be explained by how well the true variance structure is approximated by the proxy matrix M.

Table 3. Empirical rate of rejection for a 2-level hierarchical setting fitting (i) both levels correctly, (ii) "pat" only the effect of 50 patients, (iii) "doc" only the effect of 5 doctors.

	Component 3			Component 4			Component 100		
	both	pat	doc	both	pat	doc	both	pat	doc
$h = -6$	0.95	0.95	0.88	0.93	0.93	0.83	0.90	0.89	0.82
$h = -4$	0.73	0.72	0.62	0.66	0.67	0.54	0.60	0.60	0.48
$h = -2$	0.36	0.36	0.30	0.26	0.26	0.19	0.20	0.18	0.18
$h = 0$	0.08	0.08	0.08	0.06	0.05	0.05	0.05	0.06	0.04
$h = 2$	0.06	0.06	0.06	0.14	0.15	0.13	0.22	0.21	0.17
$h = 4$	0.28	0.29	0.21	0.49	0.49	0.38	0.59	0.59	0.48
$h = 6$	0.64	0.64	0.52	0.84	0.85	0.73	0.89	0.89	0.80

NOTE: There are 3 repeated measurements per patient, $p = 500$ and $s = 5$.

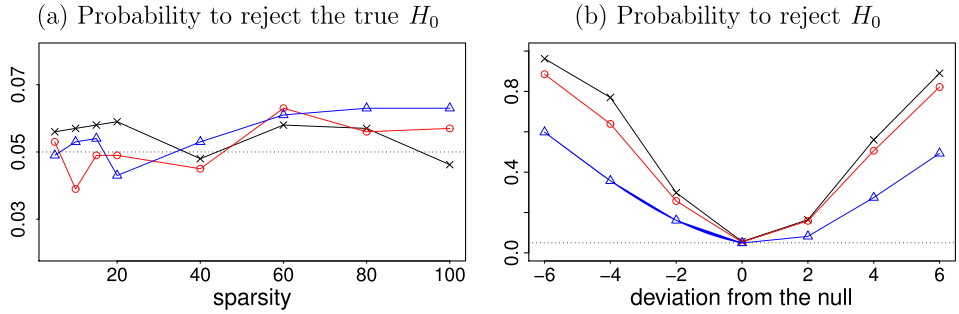

Figure 3. Empirical rate of rejection of the null hypothesis under (a) the null hypothesis, and (b) alternative hypothesis, at a 5% nominal level for different distributions of the error: Gaussian (Model 1, black crosses), Student with 10 degrees of freedom (Model 4, red circles), and Student with 3 degrees of freedom (Model 5, blue triangles). The hypothesis test is for $H_{0,4} : \beta^*_4 = \beta_{0,4}$ versus $H_{1,4} : \beta^*_4 = \beta_{0,4} + hn^{-1/2}$.

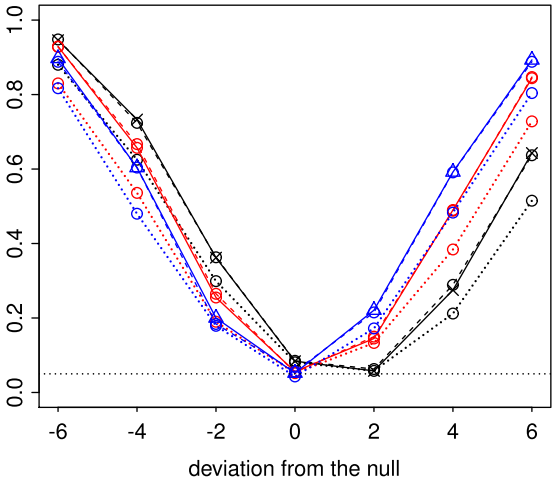

Figure 4. Empirical rate of rejection at a 5% nominal level for the hierarchical "patient-within-doctor" mixed model. Hypothesis tests are performed on different components of β^*: the third component (black crosses), the fourth component (red circles), and the 100th component (blue triangles). Results are for the correct 2-level random effect structure (solid lines), only including the patient effect (dashed) and only including the doctor effect (dotted).

Table 4. Probability to reject the null hypothesis in function of the deviation from the null for a 5% nominal level, and in function of the choice of the proxy matrix.

	Probability to reject the null hypothesis at a 5% level			
	Model 1		Model 2	
	$\boldsymbol{M} = \sigma^{-2}_{\epsilon,\text{init}}\boldsymbol{\psi}_{\text{init}}$	$\boldsymbol{M} = \log n\boldsymbol{I}_{Nq}$	$\boldsymbol{M} = \sigma^{-2}_{\epsilon,\text{init}}\boldsymbol{\psi}_{\text{init}}$	$\boldsymbol{M} = \log n\boldsymbol{I}_{Nq}$
$h = -6$	0.96	0.91	0.65	0.79
$h = -4$	0.77	0.66	0.39	0.49
$h = -2$	0.30	0.25	0.16	0.19
$h = 0$	0.06	0.05	0.05	0.05
$h = 2$	0.16	0.15	0.05	0.09
$h = 4$	0.56	0.49	0.18	0.30
$h = 6$	0.89	0.85	0.42	0.64

NOTE: Parameters are $n = 200$, $p = 500$, $N = 50$, $s = 5$. The 4th component of $\boldsymbol{\beta}^*$ is tested. Model 1 contains 2 random effects and Model 2 contains 3 random effects with large magnitude.

In our simulation setting for Model 1 the constant $\log(n)$ gives a worse approximation to the value of the diagonal entries in $\boldsymbol{\Psi}$ than when using $2/(3q)$, which results in slightly higher power for the latter choice. For Model 2, the true values of $\boldsymbol{\Psi}$ are in between $\log(n)$ and $2/(3q)$, though closer to $\log(n)$, which might explain the slightly higher power for this choice of $\boldsymbol{M}$.

4.3.5. Sensitivity to Tuning Parameters Choice

We conclude this simulation study by an analysis of the sensitivity of our procedure to the choice of the tuning parameters. We consider Model 1 and two scenarios: $(h, s) = (0, 40)$ and $(h, s) = (4, 5)$ corresponding to a nonsparse model under H_0 and an sparse model under the true alternative H_1. We perform hypothesis tests for the fourth component of $\boldsymbol{\beta}$ and provide results for different choices of the tuning parameters. Results of Table 5 illustrate that the method is reasonably weakly sensitive to the tuning parameters choice.

4.4. Hypothesis Testing for Riboflavin Data

We study the Riboflavin data which contains 4088 gene expressions from 111 observations divided in 28 groups from size 2 to 6, and which has also been considered by Schelldorfer, Buhlmann, and van de Geer (2011). The response variable is the logarithm of the riboflavin production rate of *Bacillus subtilis*. In the context of linear models a related dataset was considered in Van de Geer et al. (2014) and Javanmard and Montanari (2014b). The data of our interest have a clear group structure, and therefore we include a random intercept in the linear model, resulting into the following mixed model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{W}\boldsymbol{b} + \boldsymbol{\epsilon},$$

where W contains 1 in the appropriates entries, $\boldsymbol{Y}$ is of length 111 and $\boldsymbol{X}$ is the 111×4089 design matrix with its first column containing only ones and its other columns corresponding to the 4088 standardized covariates. The vector $\boldsymbol{b}$ contains the 28 realizations of the random effects. Its components are iid with mean zero and unknown variance $\boldsymbol{\psi}$.

Riboflavin (vitamin B2) is an essential component of the basic metabolism. The riboflavin biosynthesis in bacteria was analyzed from a biological perspective using comparative analysis of genes, operons, and regulatory elements. However, little is known about the mechanisms of regulation of the bacterial riboflavin genes.

Gene YpaA has been verified experimentally to be a transporter of riboflavin or related compounds, co-regulated with other riboflavin genes (Vitreschak et al. 2002). The RFN element (specific locus related to RNA folding) was only encoded upstream of the YpaA gene. Moreover, in *B. subtilis*, riboflavin uptake was increased when YpaA was over-expressed and abolished when YpaA was deleted (Vogl et al. 2007). Hence it makes sense to test a one sided hypothesis with the null being that the corresponding β^* is smaller or equal to zero and alternative that it is larger than zero. In this context, we have applied our test above while including all of the 4087 remaining genes in the study. We have obtained a test statistic value of 1.60. Therefore, our method is able to identify YpaA gene at a 6% significance level; we should note that our sample size is extremely small and deviations from 5% should be expected due to an extremely small sample size. It is also worth noting, that no previous

Table 5. Probability to reject the null hypothesis for different choices of the tuning parameters and for Model 1 with $j = 4$ and $(h, s) = (0, 40)$ or $(h, s) = (4, 5)$, compared with the default choice presented in Section 4.2.

	Probability to reject H_0 at a 5% level					
	$h = 0$ and $s = 40$			$h = 4$ and $s = 5$		
	$\frac{1}{2}\eta$	η	2η	$\frac{1}{2}\eta$	η	2η
0.5μ	0.06	0.05	0.05	0.48	0.48	0.53
μ	0.04	0.05	0.05	0.59	0.59	0.59
2μ	0.06	0.05	0.06	0.35	0.34	0.34

	Probability to reject H_0 at a 5% level	
	$h = 0$ and $s = 40$	$h = 4$ and $s = 5$
$\frac{1}{2}\bar{\eta}$	0.04	0.60
$\bar{\eta}$	0.05	0.59
$2\bar{\eta}$	0.04	0.42

NOTE: Notation η refers to η_γ, η_θ and η'_θ; μ refers to μ_γ and μ_θ; $\bar{\eta}$ refers to $\bar{\eta}_\gamma$ and $\bar{\eta}_\theta$. If it is not specified, the default choice is used.

study of this dataset was able to identify YpaA as statistically significant gene. We also have observed large correlation among the genes present in this study, which may additionally explain why previous methods failed; observe that our method is doubly robust and is able to overcome high correlations in the data.

The best studied system of the riboflavin biosynthesis in bacteria is the *rib* operon of *B. subtilis*. Our data contain information about the *B. subtilis* and there are a number of genes (but not all) related to *rib* operon. We are particularly interested in the ribB gene which was reported as over-expressed in a number of chemical and biological studies (Mörtl et al. 1996) but has not been yet studied from a statistical point of view. We have performed a single hypothesis test related to the β^* corresponding to ribB gene; while including all of the remaining genes. We have found that our test statistic is able to detect ribB with a 1% nominal value; the observed value of the test statistics is 2.69. This result confirms the biological evidence and reconfirms wide-applicability of the proposed test.

Another part of the *rib* operon is a ribC gene. The ribC gene was cloned and sequenced and it was determined that it plays an essential role in the flavin metabolism of *B. subtilis* (Mack, van Loon, and Hohmann 1998). In particular it was observed that it is over-expressed and that it suppresses the riboflavin overproduction. We have performed a one-sided hypothesis test to observe statistical significance of this gene in the current dataset. We have observed a test statistic value of 1.01 leading to a 16% significance level. Although larger than 10%, this study is one of the first statistical studies that was able to detect ribC gene at any level. For completeness, we also report that an averaged size of the selected sets in the original and the feature model was 17.90 and 741.10.

5. Further Explorations

The methodology presented in the previous sections is extremely general and provides easy extensions to a number of interesting problems and settings.

5.1. Power Improvements

Our test statistic uses a proxy matrix $\tilde{\boldsymbol{P}}$ which has some influence on the power through the deviation term $h\sigma_u\sigma_\epsilon^{-1}\frac{n^{-1}\text{trace}(\tilde{\boldsymbol{P}})}{\sqrt{n^{-1}\text{trace}(\tilde{\boldsymbol{P}}\boldsymbol{P}^{-1}\tilde{\boldsymbol{P}})}}$ in Theorem 4. This deviation term is smaller than $\sqrt{n^{-1}\text{trace}(\boldsymbol{P})}$ and it is expected that the "closer" $\tilde{\boldsymbol{P}}$ is to $\boldsymbol{P}$, the bigger this deviation term is and thus the better the power of the test statistics is. By using a good estimator $\hat{\boldsymbol{P}}$ of $\boldsymbol{P}$, we could expect the deviation term to be close to $h\sigma_u\sigma_\epsilon^{-1}\sqrt{n^{-1}\text{trace}(\boldsymbol{P})}$.

This leads to the following scheme:

Step 1: Compute an estimator $\hat{\boldsymbol{\gamma}}$ by solving (10) with the proxy matrix $\tilde{\boldsymbol{P}}$. We know by Lemma B.6 (see the supplementary materials) that $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_1$ is small.

Step 2: Obtain a better estimator $\hat{\boldsymbol{P}}$ of $\boldsymbol{P}$ by solving (17) .

Step 3: Compute a new estimator of $\boldsymbol{\gamma}^*$ obtained by solving (10) with $\hat{\boldsymbol{P}}$ instead of $\tilde{\boldsymbol{P}}$ and construct a T_n statistic based on this new estimator and on $\hat{\boldsymbol{P}}$.

As maximum likelihood estimators are often biased for the estimation of the variance components, we utilized the marginal log-likelihood and propose to consider the following restricted maximum likelihood estimator

$$\hat{\boldsymbol{P}} = \arg\min_{\boldsymbol{P}\geq 0}\left\{\frac{1}{2}(\boldsymbol{V}-\boldsymbol{X}\hat{\boldsymbol{\gamma}})^\top\boldsymbol{P}(\boldsymbol{V}-\boldsymbol{X}\hat{\boldsymbol{\gamma}}) + \frac{1}{2}\log(\det(\boldsymbol{P}^{-1})) + \frac{1}{2}\log(\det(\boldsymbol{X}^\top\boldsymbol{P}\boldsymbol{X}))\right\}. \tag{17}$$

Theorems 3 and 4 are expected to hold with this new T_n statistic. Implementation of this approach is nontrivial and is out of the scope of the current work; some solutions can be found in Tan et al. (2018).

With this approach, not only do we obtain a better test statistic (more efficient) but we are able to estimate the error variance of the initial model by a new estimator which can be defined as $\hat{\sigma}_\epsilon^2 = n\hat{\sigma}^2/\text{trace}(\hat{\boldsymbol{P}})$.

5.2. Multivariate Testing

For testing general multivariate hypotheses of the kind $H_0 : \boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ versus $H_1 : \boldsymbol{\beta}^* \neq \boldsymbol{\beta}_0$ where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ and $d \to \infty$ ($p \to \infty$) we can easily adapt the procedure of Section 2.

With a little abuse of notation let $\boldsymbol{V} = \boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}_0$ denote the pseudo-response vector (similar to previous sections) where now a univariate parameter β_0 is replaced with a d-dimensional counterpart $\boldsymbol{\beta}_0$. We denote by $\boldsymbol{Z}_{(j)}$ the jth column of $\boldsymbol{Z}$, $j = 1,\ldots,d$, by $\tilde{\boldsymbol{P}}_k$ the kth row of $\tilde{\boldsymbol{P}}$, and with a slight abuse in notation, we denote with $\boldsymbol{Z}_{k,(j)}$, the kth element of the vector $\boldsymbol{Z}_{(j)}$, $k = 1,\ldots,n$. Then with $\hat{\boldsymbol{\gamma}}$ as defined before, we consider d estimators $\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}, \ldots, \hat{\boldsymbol{\theta}}_{(d)}$ defined as follows

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{(j)} \quad \in \quad & \arg\min_{\boldsymbol{\theta}_{(j)}\in\mathbb{R}^{p-d}} ||\boldsymbol{\theta}_{(j)}||_1 \\ \text{such that} \quad & ||n^{-1}\boldsymbol{X}^\top(\boldsymbol{Z}_{(j)} - \boldsymbol{X}\boldsymbol{\theta}_{(j)})||_\infty \leq \eta_{\theta,j} \\ & n^{-1}\boldsymbol{Z}_{(j)}^\top(\boldsymbol{Z}_{(j)} - \boldsymbol{X}\boldsymbol{\theta}_{(j)}) \geq \bar{\eta}_{\theta,j} \\ & ||\boldsymbol{Z}_{(j)} - \boldsymbol{X}\boldsymbol{\theta}_{(j)}||_\infty \leq \mu_{\theta,j} \\ & ||n^{-1}\boldsymbol{X}^\top\tilde{\boldsymbol{P}}(\boldsymbol{Z}_{(j)} - \boldsymbol{X}\boldsymbol{\theta}_{(j)})||_\infty \leq \eta_{\theta,j}\,. \end{aligned} \tag{18}$$

Then, we consider to reject the null hypothesis whenever $T_n = \max_j T_{n,j}$ is larger than the bootstrap quantile $q_{1-\alpha}$ to be defined below. The test statistics $T_{n,j}$ are defined in the same spirit of the previous section and take the form of

$$T_{n,j} = n^{-1/2}\sum_{k=1}^n T_{kj}, \quad T_{kj} = \frac{(\boldsymbol{Z}_{k,(j)} - \boldsymbol{X}_k\hat{\boldsymbol{\theta}}_{(j)})\tilde{\boldsymbol{P}}_k\left(\boldsymbol{V}-\boldsymbol{X}\hat{\boldsymbol{\gamma}}\right)}{\hat{\sigma}_{u,j}\hat{\sigma}_\epsilon},$$

where $\hat{\sigma}_\varepsilon^2 = n^{-1}||\tilde{\boldsymbol{P}}\left(\boldsymbol{V}-\boldsymbol{X}\hat{\boldsymbol{\gamma}}\right)||_2^2$, $\hat{\sigma}_{u,j}^2 = n^{-1}||\boldsymbol{Z}_{(j)} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_{(j)}||_2^2$. Note that the test statistics have mean zero under the null hypothesis.

The quantile $q_{1-\alpha}$ is defined as a $1-\alpha$ quantile of the distribution of a bootstrapped test statistic $\tilde{T}_n = \max_j \tilde{T}_{n,j}$ with

$$\tilde{T}_{n,j} = n^{-1/2}\sum_{k=1}^n \xi_k(T_{kj} - n^{-1/2}T_{n,j})$$

for a class of multipliers $\{\xi_k\}_{k=1}^n$ that are drawn from a standard Gaussian distribution, independently from the data.

For a more general hypothesis $H_{0,j} : \beta_j^* = \beta_{0,j}$ for $j \in \mathcal{J}$, $\mathcal{J} \subset \{1, 2, \ldots, p\}$ we can consider the maximum as the test statistic and denote by $T_n(\beta_{0,j})$ the test statistic as in (13) where one element of β^* is hold-out and the remaining ones are stacked in the vector γ^*. Then, we use $G_{\mathcal{J}}(c) = P\Big[\max_{j\in\mathcal{J}}(|\tilde{T}_{n,j}|) \leq c\Big]$, Then, the p-value for $H_{0,\mathcal{J}} : \beta^* = \beta_0$, against the alternative being the complement, is defined as $P_{\mathcal{J}} = 1 - G_{\mathcal{J}}(\max_{j\in\mathcal{J}} |T_n(\beta_{0,j})|)$.

5.3. Generalized Linear Mixed Models

We note that inference in generalized linear models is an extremely difficult problem, even in low-dimensional setting. Typical approaches are hindered by a difficult numerical integration and an often nonanalytic expression of a likelihood or profile likelihood function. In this subsection, we illustrate how the methodology introduced in Section 2 can be easily extended for this difficult setting. Robustness of our approach provides flexibility regarding specifying the likelihood exactly.

Let $\boldsymbol{Y}$ be the observed data vector and, conditional on the random effects, $\boldsymbol{b}$, assume that the elements of $\boldsymbol{Y}$ are independent and drawn from a distribution in the exponential family (which, for simplicity of exposition, we take with the canonical link). To complete the specification, assume a distribution for $\boldsymbol{b}$ to be dependent on variance parameters, $\boldsymbol{D}$

$$f_{\mathbf{y}_i|\boldsymbol{b}}(\mathbf{y}|\boldsymbol{b}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*, \phi) = \exp\left\{\frac{\mathbf{y}\boldsymbol{\eta}_i - c(\boldsymbol{\eta}_i)}{a(\phi)} + d(\mathbf{y}, \phi)\right\}, \quad (19)$$

where $\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\gamma}^* + \boldsymbol{Z}_i\boldsymbol{\beta}^* + \mathbf{W}_i\boldsymbol{b}_i$. With g denoting the link function, we have $b' = g^{-1}$

$$E[\mathbf{y}_i|\boldsymbol{b}] = b'(\boldsymbol{X}_i\boldsymbol{\gamma}^* + \boldsymbol{Z}_i\boldsymbol{\beta}^* + \mathbf{W}_i\boldsymbol{b}_i).$$

We propose the following test statistic for use in generalized linear mixed models

$$T_{n,j} = \frac{n^{-1/2}(\boldsymbol{Z}_{(j)} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_{(j)})^\top \tilde{\boldsymbol{P}}\left(\boldsymbol{Y} - b'(\boldsymbol{X}\hat{\boldsymbol{\gamma}} + \boldsymbol{Z}\boldsymbol{\beta}_0)\right)}{\hat{\sigma}_{u,j}\hat{\sigma}_\epsilon}$$

with $\hat{\sigma}_\varepsilon^2 = n^{-1}||\tilde{\boldsymbol{P}}\left(\boldsymbol{Y} - b'(\boldsymbol{X}\hat{\boldsymbol{\gamma}} + \boldsymbol{Z}\boldsymbol{\beta}_0)\right)||_2^2$, $\hat{\sigma}_{u,j}^2 = n^{-1}b''(\boldsymbol{X}\hat{\boldsymbol{\gamma}} + \boldsymbol{Z}\boldsymbol{\beta}_0)||\boldsymbol{Z}_{(j)} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_{(j)}||_2^2$.

For the procedure to be adaptive to generalized linear models, the estimators of $\boldsymbol{\gamma}^*$ and $\boldsymbol{\theta}_{(j)}^*$ need to be carefully developed. Regarding the estimation of $\boldsymbol{\gamma}^*$ we adapt the estimator of Section 2.2 and propose the following estimator

$$\begin{aligned} \hat{\boldsymbol{\gamma}} \in & \quad \underset{\boldsymbol{\gamma}\in\mathbb{R}^{p-d}}{\arg\min}||\boldsymbol{\gamma}||_1 \\ \text{such that } & ||n^{-1}\boldsymbol{X}^\top\tilde{\boldsymbol{P}}\Big(\boldsymbol{Y} - b'(\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{\beta}_0)\Big)||_\infty \leq \eta_\gamma \\ & n^{-1}\boldsymbol{Y}^\top\tilde{\boldsymbol{P}}\Big(\boldsymbol{Y} - b'(\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{\beta}_0)\Big) \geq \bar{\eta}_\gamma \\ & ||\tilde{\boldsymbol{P}}(\boldsymbol{Y} - b'(\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{\beta}_0))||_\infty \leq \mu_\gamma \end{aligned} \quad (20)$$

for suitable choices of tuning parameters $\eta_\gamma \asymp \sqrt{n^{-1}\log(p)}$, $0 < \bar{\eta}_\gamma < \sigma_\varepsilon^2$ and $\mu_\gamma \asymp \sqrt{\log(n)}$. The estimator for $\boldsymbol{\theta}_{(j)}^*$ can now be defined with

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{(j)} \in & \quad \underset{\boldsymbol{\theta}_{(j)}\in\mathbb{R}^{p-d}}{\arg\min}||\boldsymbol{\theta}_{(j)}||_1 \\ \text{such that } & ||n^{-1}b''(\boldsymbol{X}\hat{\boldsymbol{\gamma}} + \boldsymbol{Z}\boldsymbol{\beta}_0)\boldsymbol{X}^\top(\boldsymbol{Z}_{(j)} - \boldsymbol{X}\boldsymbol{\theta}_{(j)})||_\infty \leq \eta_{\theta,j} \\ & ||n^{-1}b''(\boldsymbol{X}\hat{\boldsymbol{\gamma}} + \boldsymbol{Z}\boldsymbol{\beta}_0)\boldsymbol{X}^\top\tilde{\boldsymbol{P}}(\boldsymbol{Z}_{(j)} - \boldsymbol{X}\boldsymbol{\theta}_{(j)})||_\infty \leq \bar{\eta}_{\theta,j} \\ & n^{-1}b''(\boldsymbol{X}\hat{\boldsymbol{\gamma}} + \boldsymbol{Z}\boldsymbol{\beta}_0)\boldsymbol{Z}_{(j)}^\top(\boldsymbol{Z}_{(j)} - \boldsymbol{X}\boldsymbol{\theta}_{(j)}) \geq \bar{\eta}_{\theta,j} \\ & ||\boldsymbol{Z}_{(j)} - \boldsymbol{X}\boldsymbol{\theta}_{(j)}||_\infty \leq \mu_{\theta,j}, \end{aligned} \quad (21)$$

for suitable choices of tuning parameters $\eta_{\theta,j} \asymp \sqrt{n^{-1}\log(p)}$, $0 < \bar{\eta}_{\theta,j} < \sigma_{u,j}^2$, and $\mu_{\theta,j} \asymp \sqrt{\log(n)}$. Here, $\boldsymbol{Z}_{(j)} = (z_{1,(j)}, \ldots, z_{n,(j)})^\top \in \mathbb{R}^n$. Observe that differently from the linear mixed models, in the case of the generalized linear mixed models, the two estimators $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\theta}}_{(j)}$ are dependent. The above procedure can be solved using iterations of linear programs, much in the spirit of weighted least squares methods. The final test statistic is now defined in a similar manner as before with $T_n = \max_j T_{n,j}$ where the multiplier bootstrap of the previous subsection can be successfully applied.

6. Discussion

This article proposed a class of test statistics for performing inference on fixed effects that allow for high-dimensional and misspecified linear mixed models all while maintaining the benefits of classical methods, that is, an asymptotically normal and unbiased test statistic with valid and honest confidence intervals. Our test statistic can be thought of as a doubly robust approach; it adapts to the sparsity of the nuisance parameters (fixed) as well as the unknown structure of the random effects (both variance and distributions). Such adaptivity seems essential for modern large-scale applications with many features, various sparsity assumptions as well as distributional assumptions that cannot be checked.

In general, the challenge in using adaptive methods as the basis for valid statistical inference is that selection bias can be difficult to quantify. In this article, pairing the initial model with a complementary feature model enabled us to accomplish this goal in a simple yet principled way. In our simulation experiments, our method provides better error control while achieving nominal coverage rates in moderate sample sizes.

A number of important extensions and refinements are left open. Our current results only provide point-wise confidence intervals; extending our theory to the setting of multivariate testing or the setting of a generalized linear mixed models, seems like a promising avenue for further work. Another challenge is the selection of the proxy matrix $\boldsymbol{M}$ toward better efficiency and power. A systematic approach to design optimization and theory for such setting, would improve the finite sample performance. In general, work can be done to identify methods that furthermore allow for inference on the variance components and tests of heterogeneity even in more challenging circumstances, for example, with small samples or a large number of covariates are likely to be bring impactful work to a broader scientific audience.

Supplementary Materials

The supplemental material contains details to the technical results of the main document. In particular it provides detailed proofs of Theorems 1-4 and establishes a sequence of useful Lemmas with proofs.

Acknowledgments

The authors thank the reviewers for their comments that helped improve the article.

Funding

Gerda Claeskens and Thomas Gueuning would like to acknowledge the support of the Research Foundation Flanders and KU Leuven grant GOA/12/14. Jelena Bradic would like to acknowledge the support of the National Science Foundation grant DMS-1712481. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government—Department EWI.

References

Anderson, T. W., ed. (1984), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley. [1838]

Athey, S., Imbens, G. W., and Wager, S. (2016), "Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions," arXiv no. 1604.07125. [1836]

Belloni, A., Chernozhukov, V., Chetverikov, D., and Wei, Y. (2015), "Uniformly Valid Post-Regularization Confidence Regions for Many Functional Parameters in z-Estimation Framework," arXiv no. 1512.07619. [1836]

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017), "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85, 233–298. [1836]

Belloni, A., Chernozhukov, V., and Kato, K. (2014), "Uniform Post-Selection Inference for Least Absolute Deviation Regression and Other z-Estimation Problems," *Biometrika*, 102, 77–94. [1836]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [1841]

Bonnet, A., Gassiat, E., and Lévy-Leduc, C. (2015), "Heritability Estimation in High Dimensional Sparse Linear Mixed Models," *Electronic Journal of Statistics*, 9, 2099–2129. [1837]

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25. [1835,1841]

Bühlmann, P., and Van De Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Berlin, Heidelberg: Springer Science & Business Media. [1836]

Cai, T. T., and Guo, Z. (2017), "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," *The Annals of Statistics*, 45, 615–646. [1836]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. K. (2016), "Double Machine Learning for Treatment and Causal Parameters," arXiv no. 1608.00060. [1836]

Chernozhukov, V., Hansen, C., and Spindler, M. (2015), "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics*, 7, 649–688. [1836]

Crainiceanu, C. M., and Ruppert, D. (2004), "Likelihood Ratio Tests in Linear Mixed Models With One Variance Component," *Journal of the Royal Statistical Society*, Series B, 66, 165–185. [1835]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American statistical Association*, 96, 1348–1360. [1836]

Fan, Y., and Li, R. (2012), "Variable Selection in Linear Mixed Effects Models," *The Annals of Statistics*, 40, 2043–2068. [1837,1839,1841,1844,1845]

Ghosh, A., and Thoresen, M. (2016), "Non-Concave Penalization in Linear Mixed-Effects Models and Regularized Selection of Fixed Effects," arXiv no. 1607.02883. [1837]

Goeman, J. J., Van Houwelingen, H. C., and Finos, L. (2011), "Testing Against a High-Dimensional Alternative in the Generalized Linear Model: Asymptotic Type I Error Control," *Biometrika*, 98, 381–390. [1841]

Groll, A., and Tutz, G. (2014), "Variable Selection for Generalized Linear Mixed Models by L_1-Penalized Estimation," *Statistics and Computing*, 24, 137–154. [1837]

Heagerty, P. J., and Kurland, B. F. (2001), "Misspecified Maximum Likelihood Estimates and Generalised Linear Mixed Models," *Biometrika*, 88, 973. [1842]

Hobert, J. P., and Casella, G. (1996), "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models," *Journal of the American Statistical Association*, 91, 1461–1473. [1842]

Hui, F. K., Müller, S., and Welsh, A. (2017), "Joint Selection in Mixed Models Using Regularized PQL," *Journal of the American Statistical Association*, 112, 1323–1333. [1837]

Jankova, J., and Van De Geer, S. (2015), "Confidence Intervals for High-Dimensional Inverse Covariance Estimation," *Electronic Journal of Statistics*, 9, 1205–1229. [1836]

Javanmard, A., and Montanari, A. (2014a), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [1836]

——— (2014b), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [1842,1846]

Kenward, M. G., and Roger, J. H. (1997), "Small Sample Inference for Fixed Effects From Restricted Maximum Likelihood," *Biometrics*, 53, 983–997. [1835]

Koenker, R., and Mizera, I. (2014), "Convex Optimization in R," *Journal of Statistical Software*, 60(5), 1–23. [1840]

Lachos, V. H., Ghosh, P., and Arellano-Valle, R. B. (2010), "Likelihood Based Inference for Skew-Normal Independent Linear Mixed Models," *Statistica Sinica*, 20, 303–322. [1841]

Lindstrom, M. J., and Bates, D. M. (1988), "Newton–Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, 83, 1014–1022. [1841]

Litiére, S., Alonso, A., and Molenberghs, G. (2007), "Type I and Type II Error Under Random-Effects Misspecification in Generalized Linear Mixed Models," *Biometrics*, 63, 1038–1044. [1842]

Mack, M., van Loon, A. P., and Hohmann, H.-P. (1998), "Regulation of Riboflavin Biosynthesis in *Bacillus subtilis* Is Affected by the Activity of the Flavokinase/Flavin Adenine Dinucleotide Synthetase Encoded by*ribC*," *Journal of Bacteriology*, 180, 950–955. [1847]

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170. [1841]

McGilchrist, C. (1994), "Estimation in Generalized Mixed Models," *Journal of the Royal Statistical Society*, Series B, 56, 61–69. [1841]

Mörtl, S., Fischer, M., Richter, G., Tack, J., Weinkauf, S., and Bacher, A. (1996), "Biosynthesis of Riboflavin Lumazine Synthase of *Escherichia coli*," *Journal of Biological Chemistry*, 271, 33201–33207. [1847]

Müller, S., Scealy, J., and Welsh, A. (2013), "Model Selection in Linear Mixed Models," *Statistical Science*, 28, 135–167. [1837]

Neyman, J. (1959), "Optimal Asymptotic Tests of Composite Statistical Hypotheses," *Probability and Statistics*, 57, 213. [1836]

Ning, Y., and Liu, H. (2017), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," *The Annals of Statistics*, 45, 158–195. [1836]

Quintana, F. A., Johnson, W. O., Waetjen, L. E., and Gold, E. B. (2016), "Bayesian Nonparametric Longitudinal Data Analysis," *Journal of the American Statistical Association*, 111, 1168–1181. [1837]

Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. (2015), "Asymptotic Normality and Optimalities in Estimation of Large Gaussian Graphical Models," *The Annals of Statistics*, 43, 991–1026. [1836]

Rohart, F., San-Cristobal, M., and Laurent, B. (2014), "Fixed Effects Selection in High Dimensional Linear Mixed Models, Using a Multicycle ECM Algorithm." *Computational Statistics and Data Analysis*, 80, 209-222. [1843,1844,1845]

Rudelson, M., and Zhou, S. (2013), "Reconstruction From Anisotropic Random Measurements," *IEEE Transactions on Information Theory*, 59, 3434–3447. [1840,1841]

Ryzhov, I. O., Han, B., and Bradic, J. (2016), "Cultivating Disaster Donors Using Data Analytics," *Management Science*, 62, 849–866. [1836]

Schelldorfer, J., Buhlmann, P., and van de Geer, S. (2011), "Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization," *Scandinavian Journal of Statistics*, 38, 197–214. [1837,1839,1841,1842,1843,1846]

Song, P. X.-K., Zhang, P., and Qu, A. (2007), "Maximum Likelihood Inference in Robust Linear Mixed-Effects Models Using Multivariate t Distributions," *Statistica Sinica*, 17, 929–943. [1841]

Tan, Z., Roche, K., Zhou, X., and Mukherjee, S. (2018), "Scalable Algorithms for Learning High-Dimensional Linear Mixed Models," arXiv no. 1803.04431. [1847]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1836]

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [1836, 1842,1846]

Verbeke, G., and Lesaffre, E. (1996), "A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population," *Journal of the American Statistical Association*, 91, 217–221. [1841]

Verbeke, G., and Molenberghs, G. (2009), *Linear Mixed Models for Longitudinal Data* (corrected edition), New York: Springer. [1835]

Vitreschak, A. G., Rodionov, D. A., Mironov, A. A., and Gelfand, M. S. (2002), "Regulation of Riboflavin Biosynthesis and Transport Genes in Bacteria by Transcriptional and Translational Attenuation," *Nucleic Acids Research*, 30, 3141–3151. [1846]

Vogl, C., Grill, S., Schilling, O., Stülke, J., Mack, M., and Stolz, J. (2007), "Characterization of Riboflavin (Vitamin B2) Transport Proteins From *Bacillus subtilis* and *Corynebacterium glutamicum*," *Journal of Bacteriology*, 189, 7367–7375. [1846]

Wang, L., Zhou, J., and Qu, A. (2012), "Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis," *Biometrics*, 68, 353–360. [1837]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [1836]

Zhang, D., and Davidian, M. (2001), "Linear Mixed Models With Flexible Distributions of Random Effects for Longitudinal Data," *Biometrics*, 57, 795–802. [1841]

Zhu, Y., and Bradic, J. (2018), "Linear Hypothesis Testing in Dense High-Dimensional Linear Models," *Journal of the American Statistical Association*, 113, 1583–1600. [1843,1844,1845]