Journal of the American Statistical Association

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

Inference and Estimation for Random Effects in High-Dimensional Linear Mixed Models

Michael Law & Ya'acov Ritov

To cite this article: Michael Law & Ya'acov Ritov (2023) Inference and Estimation for Random Effects in High-Dimensional Linear Mixed Models, Journal of the American Statistical Association, 118:543, 1682-1691, DOI: 10.1080/01621459.2021.2004896

To link to this article: https://doi.org/10.1080/01621459.2021.2004896

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 543, 1682–1691: Theory and Methods
https://doi.org/10.1080/01621459.2021.2004896

Inference and Estimation for Random Effects in High-Dimensional Linear Mixed Models

Michael Law and Ya'acov Ritov

Department of Statistics, University of Michigan, Ann Arbor, MI

ABSTRACT

We consider three problems in high-dimensional linear mixed models. Without any assumptions on the design for the fixed effects, we construct asymptotic statistics for testing whether a collection of random effects is zero, derive an asymptotic confidence interval for a single random effect at the parametric rate $\sqrt{n}$, and propose an empirical Bayes estimator for a part of the mean vector in ANOVA type models that performs asymptotically as well as the oracle Bayes estimator. We support our theoretical results with numerical simulations and provide comparisons with oracle estimators. The procedures developed are applied to the Trends in International Mathematics and Sciences Study (TIMSS) data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2020
Accepted October 2021

KEYWORDS
Exponential weights; High-dimensional; Random effects; Variance components

1. Introduction

In the past two decades, there has been a lot of progress in the theory for high-dimensional linear models. However, its close cousin, the high-dimensional linear mixed model, has received significantly less attention; it was not until the past decade until there were procedures for estimation. Consider a linear mixed model given by

$$Y = \mu + Z\nu + W\gamma + \varepsilon, \tag{1}$$

with $Z \in \mathbb{R}^{n\times q}$, $W \in \mathbb{R}^{n\times d}$, and $Y, \mu, \varepsilon \in \mathbb{R}^n$; the vector μ and the pair ν and γ are the fixed effects and the random effects, respectively. In addition, we observe covariates $X \in \mathbb{R}^{n\times p}$ such that $\mu \approx X\beta$ for some sparse vector $\beta \in \mathbb{R}^p$ (see Section 1.2 for a rigorous definition). Here, X is the component of the design corresponding to the fixed effects and (Z, W) the component corresponding to the random effects. We consider the setting where the random effects are low-dimensional, $q + d < n$, but the fixed effects are high-dimensional, $p > n$. We have separated the random effects in two to emphasize that later we are interested in ν and view γ as nuisance parameters. Various authors have considered different aspects of this problem.

The earliest work of Schelldorfer, Bühlmann, and van de Geer (2011) proposed an estimator for both β and the variance components using a lasso-type approach. These types of approaches were later extended by several authors who considered estimation with both convex penalties, such as Groll and Tutz (2014), and nonconvex penalties, such as Wang, Zhou, and Qu (2012). There is also a growing literature on model selection in high-dimensional linear mixed models (see, e.g., the review article by Müller, Scealy, and Welsh 2013).

The problem of inference is slightly less well studied. To the best of our knowledge, hypotheses testing problems were first considered by Chen et al. (2015) for random effects and Bradic, Claeskens, and Gueuning (2017) for fixed effects. However, the work of Chen et al. (2015) only consider the special case of ANOVA designs for random effects. During the preparation of this manuscript, we became aware of the independent work of Li, Cai, and Li (2021), who consider the problem of inference in high-dimensional linear mixed models. In particular, they discuss inference for fixed effects and estimation of variance components. A more detailed comparison of our methodology with Li, Cai, and Li (2021) is deferred to Section 2.4. We also note that there is a parallel notion of high-dimensional mixed models, where the number of fixed effects is low-dimensional while the random effects are high-dimensional. Under this setting, Jiang et al. (2016) established asymptotic results for the restricted maximum likelihood for variance components.

The goal of the present article is to contribute to this growing literature on high-dimensional linear mixed models where the fixed effects are high-dimensional, both in terms of estimation and inference. In particular, we consider three related problems:

1. Testing whether a collection of random effects is zero.
2. Constructing confidence intervals for the variance of a single random effect.
3. Estimating using empirical Bayes in Gaussian ANOVA Type Models.

Our methodology is inspired by both low-dimensional linear mixed models as well as high-dimensional linear models. Specifically, our approach to all three problems starts with considering a procedure in the corresponding low-dimensional problem and retrofitting it with tools and techniques from high-dimensional linear models to produce a procedure for high-dimensional linear mixed models. Throughout the article, while we consider the general linear mixed effects models, we use the balanced

CONTACT Michael Law *mmylaw@umich.edu* University of Michigan, 1085 S. University Ave. West Hall, Ann Arbor, MI 48109-1382.

one-way ANOVA model to simplify the discussion of our estimators and assumptions.

1.1. Organization of the Paper

We end the current section with a description of the notation that we adopt throughout the article. Sections 2, 3, and 4 consider the three problems outlined in the Introduction in succession. Each one starts with a description of the problem setup, a brief motivation from the low-dimensional problem, and a description of the estimator, that is, considered, and ends with some theoretical results. In Sections 5 and 6, we provide the results of our simulations and a real data application, respectively. For the ease of presentation, we defer all proofs and additional simulation results to the supplementary material.

1.2. Notation

Throughout, all of our variables have a dependence on n, but we suppress this dependence when it does not cause confusion. For a general vector a and matrix A, let $\|a\|_2$ denote the standard Euclidean norm with the dimension of the space being implicit from the vector, $\|A\|_2$ the operator norm, and $\|A\|_{\text{HS}}$ the Hilbert–Schmidt norm. Furthermore, if A is square, then $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximal and minimal eigenvalue of A, respectively. For any $k \in \mathbb{N}$, we let $\lambda_{\max,k}(A)$ denote the kth largest eigenvalue of A if A is square. Moreover, we write $1_k \in \mathbb{R}^k$ and $I_k \in \mathbb{R}^{k\times k}$ to denote the k-dimensional vector of all ones and the k-dimensional identity matrix, respectively. For two matrices A and B, the notation $A \ominus B$ denotes the intersection of the column space of A and the orthogonal complement of the column space of B. Then, for a matrix A, we write P_A to denote the projection onto the column space of A and $P_A^{\perp}$ the projection onto the orthogonal complement. Moreover, we write r_A to denote the rank of A.

Consistent with other high-dimensional works, we assume that β is a sparse vector. There are various notions of sparsity but we assume the slightly more general setting of weak sparsity from Law and Ritov (2021). Before providing the definition, we introduce some more notation. For $u \in \mathbb{N}$, we let $\mathcal{M}_u \triangleq \{m \subseteq \{1,\dots,p\} : |m| = u\}$ denote the collection of all models with the dimension of the fixed effects design equal to u. For a model $m \in \mathcal{M}_u$, X_m denotes the $n \times u$ sub-matrix of X corresponding to the columns indexed by m.

Definition 1. The vector μ is said to satisfy the *weak sparsity property relative to* X with sparsity s at rate k as $n \to \infty$ if the set

$$\mathcal{S}_\mu \triangleq \left\{ m \in \mathcal{M}_s : \|P^{\perp}_{X_m}\mu\|_2^2 = o(k) \right\}$$

is nonempty.

Then, we let $S \in \mathcal{S}_\mu$ denote any weakly sparse set for μ. We note that the usual high-dimensional setting of strong sparsity, where $\mu = X_S\beta_S$ for $|S| = s$, implies that μ is weakly sparse relative to X with sparsity s. Similar to other works on high-dimensional linear models and high-dimensional linear mixed models, we consider errors and random effects which are sub-Gaussian, for which we use the following definition:

Definition 2. A random vector $\xi \in \mathbb{R}^n$ is said to be *sub-Gaussian* with parameter K if

$$\mathbb{E}\exp\left(\lambda^{\mathsf{T}}\xi\right) \le \exp\left(\frac{K^2\|\lambda\|_2^2}{2}\right)$$

for all $\lambda \in \mathbb{R}^n$.

Note that if ξ is sub-Gaussian with parameter K and $A \in \mathbb{R}^{a\times n}$ is any deterministic matrix, then $A\xi$ is also sub-Gaussian with parameter $K\lambda_{\max}(A^{\mathsf{T}}A)$. Finally, the asymptotic distributions of some of our estimators depend on the fourth moments of the underlying distributions. We write $\kappa_\varepsilon \triangleq \text{var}(\varepsilon_1^2)$, $\omega_\varepsilon \triangleq \mathbb{E}(\varepsilon_1^4)$, $\kappa_\nu \triangleq \text{Var}(\nu_1^2)$, and $\omega_\nu \triangleq \mathbb{E}(\nu_1^4)$ when ν corresponds to a single random effect.

2. Hypotheses Testing for Random Effects

In this section, we consider the problem of inference for a collection of random effects. Consider the high-dimensional linear mixed model (1) and let $\Psi \triangleq \text{var}(\nu)$. We are interested in the hypotheses testing problem

$$H_0 : \lambda_{\max}(\Psi) = 0, \qquad H_1 : \lambda_{\max}(\Psi) > 0. \tag{2}$$

We propose two procedures in this section depending on whether ε is Gaussian.

2.1. Model and Motivation

Suppose temporarily that we are in the low-dimensional Gaussian setting with $s = p$, $p + q + d < n$, $\mu = X_S\beta_S$, $\varepsilon \sim \mathcal{N}_n(0_n, \sigma_\varepsilon^2 I_n)$ for some positive constant $\sigma_\varepsilon^2 > 0$, and $\nu \sim \mathcal{N}_q(0_q, \Psi)$ for some symmetric positive semi-definite matrix Ψ. Then, in this problem, the standard procedure for testing ν is through the Wald F-test. Writing $r_{Z\ominus(X_S,W)} \triangleq \text{rank}(P_{Z\ominus(X_S,W)})$ and $r_{(X_S,Z,W)^\perp} \triangleq \text{rank}(P^{\perp}_{(X_S,Z,W)})$, the Wald F-test is defined as

$$F_{\text{ld}} = \frac{\|P_{Z\ominus(X_S,W)}Y\|_2^2/r_{Z\ominus(X_S,W)}}{\|P^{\perp}_{(X_S,Z,W)}Y\|_2^2/r_{(X_S,Z,W)^\perp}}. \tag{3}$$

Under the null hypothesis, the above statistic has an $F_{r_{Z\ominus(X_S,W)},r_{(X_S,Z,W)^\perp}}$ distribution. The main obstacle to directly using the Wald F-test in the high-dimensional setting is removing the contribution of the fixed effects. One possibility is to perform model selection and choose the relevant covariates from X and then use the Wald F-test. Chen et al. (2015) consider a similar problem in the growing dimensional setting and they use a SCAD based approach for variable selection. As a consequence, they require $p = o(\sqrt{n})$. Instead, we leverage the fact that a projection onto a particular space is a regression onto a design whose columns span the same space.

Expanding both the numerator and the denominator of the Wald F-statistic, we have that

$$P_{Z\ominus(X_S,W)}Y = P_{Z\ominus(X_S,W)}Z\nu + P_{Z\ominus(X_S,W)}\varepsilon,$$
$$P^{\perp}_{(X,Z,W)}Y = P^{\perp}_{(X_S,Z,W)}\varepsilon.$$

In both matrices above, they project onto the orthogonal complement of W, which may still be achieved in the high-dimensional problem since W is a low-dimensional matrix.

Thus, we may find two projection matrices, $P_{Z\ominus W}$ and $P^{\perp}_{(Z,W)}$, such that

$$P_{Z\ominus W}Y = P_{Z\ominus W}X\beta + P_{Z\ominus W}Z\nu + P_{Z\ominus W}\varepsilon,$$
$$P^{\perp}_{(Z,W)}Y = P^{\perp}_{(Z,W)}X\beta + P^{\perp}_{(Z,W)}\varepsilon.$$

If $P_{Z\ominus W}X$ was low-dimensional, obtaining the projection of $P_{Z\ominus W}Y$ onto the orthogonal complement of $P_{Z\ominus W}X$ is equivalent to finding the residuals of $P_{Z\ominus W}Y$ using the covariates $P_{Z\ominus W}X$; this yields $P_{Z\ominus W}Y - P_{Z\ominus W}X\hat{\beta}$, where $\hat{\beta}$ is the least-squares estimator for β. The same holds for $P^{\perp}_{(Z,W)}X$ and $P^{\perp}_{(Z,W)}Y$. Then, we have that

$$\begin{aligned} P_{Z\ominus W}Y - P_{Z\ominus W}X\hat{\beta} &= (P_{Z\ominus W}X\beta - P_{Z\ominus W}X\hat{\beta}) \\ &\quad + P_{Z\ominus W}Z\nu + P_{Z\ominus W}\varepsilon, \\ P^{\perp}_{(Z,W)}Y - P^{\perp}_{(Z,W)}X\hat{\beta} &= (P^{\perp}_{(Z,W)}X\beta - P^{\perp}_{(Z,W)}X\hat{\beta}) + P^{\perp}_{(Z,W)}\varepsilon. \end{aligned}$$

Hence, this recasts the problem into one of high-dimensional prediction, for which there have been many procedures suggested to estimate $P_{Z\ominus W}X\beta$ and $P^{\perp}_{(Z,W)}X\beta$, such as the lasso and exponential weighting (see Tibshirani 1996 and Leung and Barron 2006, respectively). Therefore, we propose using a plug-in estimator for $P_{Z\ominus W}X\beta$ and $P^{\perp}_{(Z,W)}X\beta$ using exponential weighting of all models of a particular size and then consider the resultant residuals. Since we view the fixed effects as nuisance parameters, we consider exponential weighting instead of the lasso since exponential weighting does not require any assumptions on the design matrix X. However, most of the theory developed also applies to other plug-in estimators, albeit with simple modifications and much stronger conditions. This idea, under some mild assumptions, provides an asymptotic F-test.

However, there are two asymptotic regimes for the random effects: (i) the number of random effects increases to infinity and (ii) the number of random effects stays bounded. These two settings require slightly different analyses, so we consider separate exponential weighting estimators for the two cases.

Besides providing an asymptotic F distribution when ε is Gaussian, the F-ratio in Equation (3) simultaneously removes the scaling effect from σ_ε^2. When ε is known only to be sub-Gaussian, the ratio no longer follows an F-distribution. However, after appropriate rescaling, we may still achieve the ancillary property relative to σ_ε^2 by looking at the difference instead of the ratio. This approach, under slightly stronger sparsity assumptions, leads to an asymptotic z-test with only the sub-Gaussian assumption on the error distribution.

2.2. Estimator

In the setting, where the number of random effects increases to infinity, instead of estimating $P_{Z\ominus W}X\beta$ and $P^{\perp}_{(Z,W)}X\beta$ separately, we estimate $P^{\perp}_{W}X\beta$ and then project the resultant vector onto $P_{Z\ominus W}$ and $P^{\perp}_{(Z,W)}$, respectively. In addition to saving on computational time by only using exponential weighting once, this also allows us to leverage a larger sample size when estimating the mean vector. To apply exponential weighting, we fix a sequence of sparsities $u = u_n$. Let $\hat{\beta}_m$ denote the least-squares estimator of β using the model $m \in \mathcal{M}_u$ with covariates $P^{\perp}_{W}X_m$. Let $K_{Z\nu+\varepsilon}$ denote the sub-Gaussian parameter for $Z\nu + \varepsilon$. We define the exponential weights by

$$w_m \triangleq \frac{\exp\left(-\frac{1}{\alpha}\|P^{\perp}_{W}(Y - X\hat{\beta}_m)\|_2^2\right)}{\sum_{k\in\mathcal{M}_u}\exp\left(-\frac{1}{\alpha}\|P^{\perp}_{W}(Y - X\hat{\beta}_k)\|_2^2\right)},$$

where $\alpha > 4K_{Z\nu+\varepsilon}$. Then, the estimator for β is given by

$$\hat{\beta}_{\mathrm{EW}} \triangleq \textstyle\sum_{m\in\mathcal{M}_u} w_m\hat{\beta}_m.$$

Note that the bound on α is to ensure $P^{\perp}_{W}X\hat{\beta}_{\mathrm{EW}}$ is a consistent estimator of $P^{\perp}_{W}X\beta$. In the case where both ν and ε are Gaussian, the above bound on α becomes $\alpha > 4\left(\sigma_\varepsilon^2 + \lambda_{\max}\left(Z\Psi Z^{\mathsf{T}}\right)\right)$. Then, we estimate $P_{Z\ominus W}X\beta$ and $P^{\perp}_{(Z,W)}X\beta$ by $P_{Z\ominus W}X\hat{\beta}_{\mathrm{EW}}$ and $P^{\perp}_{(Z,W)}X\hat{\beta}_{\mathrm{EW}}$, respectively. The corresponding F-statistic is

$$F_{\mathrm{EW}} \triangleq \frac{\|P_{Z\ominus W}(Y - X\hat{\beta}_{\mathrm{EW}})\|_2^2/r_{Z\ominus W}}{\|P^{\perp}_{(Z,W)}(Y - X\hat{\beta}_{\mathrm{EW}})\|_2^2/r_{(Z,W)^\perp}}.$$

Similar to the Wald F-statistic, we reject the null hypothesis for large values of F_{EW}. In particular, for a value $\delta \in (0,1)$, let $F_{a,b,\delta}$ denote the δ upper quantile of the $F_{a,b}$ distribution. Then, we consider tests of the form

$$\phi_{F,\delta} \triangleq \mathbb{1}\left(F_{\mathrm{EW}} > F_{r_{Z\ominus W}, r_{(Z,W)^\perp},\delta}\right).$$

For the second setting where the number of random effects stay bounded, we estimate the numerator differently. Let $U_{(Z,W)^\perp} \in \mathbb{R}^{n\times r_{(Z,W)^\perp}}$ be any orthogonal matrix such that $P^{\perp}_{(Z,W)} = U_{(Z,W)^\perp}U^{\mathsf{T}}_{(Z,W)^\perp}$; for example, the matrix $U_{(Z,W)^\perp}$ may be computed by taking the spectral decomposition of $P^{\perp}_{(Z,W)}$. Define $\tilde{Y} = U^{\mathsf{T}}_{(Z,W)^\perp}Y$ and let $\tilde{Y}^{(1)}, \tilde{Y}^{(2)} \in \mathbb{R}^{r_{(Z,W)^\perp}/2}$ be a partition of $\tilde{Y}$. We similarly define $\tilde{X}^{(1)}$ and $\tilde{X}^{(2)}$. Then, letting $\tilde{\beta}^{(1)}_m$ (respectively, $\tilde{\beta}^{(2)}_m$) denote the least-squares estimator of β using the model $m \in \mathcal{M}_u$ with covariates $\tilde{X}^{(1)}_m$ (respectively, $\tilde{X}^{(2)}_m$), the exponential weights are defined as

$$\tilde{w}^{(1)}_m \triangleq \frac{\exp\left(-\frac{1}{\tilde{\alpha}}\|Y^{(1)} - X^{(1)}_m\tilde{\beta}^{(1)}_m\|_2^2\right)}{\sum_{k\in\mathcal{M}_u}\exp\left(-\frac{1}{\tilde{\alpha}}\|Y^{(1)} - X^{(1)}_k\tilde{\beta}^{(1)}_k\|_2^2\right)}$$

and similarly for $\tilde{w}^{(2)}_m$, where $\tilde{\alpha}$ is delineated in Theorem 4 below. Now, define

$$\tilde{\beta}^{(1)}_{\mathrm{EW}} \triangleq \textstyle\sum_{m\in\mathcal{M}_u}\tilde{w}^{(1)}_m\tilde{\beta}^{(2)}_m, \qquad \tilde{\beta}^{(2)}_{\mathrm{EW}} \triangleq \textstyle\sum_{m\in\mathcal{M}_u}\tilde{w}^{(2)}_m\tilde{\beta}^{(1)}_m.$$

Then, the estimator of β is

$$\tilde{\beta}_{\mathrm{EW}} \triangleq (\tilde{\beta}^{(1)}_{\mathrm{EW}} + \tilde{\beta}^{(2)}_m)/2$$

and the corresponding F-statistic is

$$\tilde{F}_{\mathrm{EW}} \triangleq \frac{\|P_{Z\ominus W}(Y - X\tilde{\beta}_{\mathrm{EW}})\|_2^2/r_{Z\ominus W}}{\|P^{\perp}_{(Z,W)}(Y - X\tilde{\beta}_{\mathrm{EW}})\|_2^2/r_{(Z,W)^\perp}}.$$

At first sight, computation of these estimators may seem prohibitive since we need to aggregate over $\binom{p}{u}$ models. However, they may be well approximated by an MCMC algorithm given by Law and Ritov (2021), to which we refer the interested reader.

In the setting, where the ε is not distributed Gaussian, we consider the following z-statistic

$$z_{\mathrm{EW}} \triangleq \|P_{Z\ominus W}(Y - X\hat{\beta}_{\mathrm{EW}})\|_2^2 - r_{Z\ominus W}r^{-1}_{(Z,W)^\perp}\|P^{\perp}_{(Z,W)}(Y - X\hat{\beta}_{\mathrm{EW}})\|_2^2.$$

Under proper scaling, the statistic z_{EW} has an asymptotic Gaussian distribution under the null hypothesis. Let

$$\sigma_{\varsigma,z}^2 \triangleq \kappa_\varepsilon \sum_{i=1}^{n} (P_{Z\ominus W} - r_{Z\ominus W} r_{(Z,W)^\perp}{}^{-1} P_{(Z,W)}^\perp)_{i,i}^2 + 2\sigma_\varepsilon^4 \sum_{i\neq j} (P_{Z\ominus W} - r_{Z\ominus W} r_{(Z,W)^\perp}{}^{-1} P_{(Z,W)}^\perp)_{i,j}^2,$$

with $\hat{\sigma}_{\varsigma,z}^2$ a consistent estimator of $\sigma_{\varsigma,z}^2$. The quantity $\sigma_{\varsigma,z}^2$ is the scaling factor to ensure a central limit for z_{EW}. Then, letting z_δ denote the δ upper quantile of the standard Gaussian distribution, we consider tests of the form

$$\phi_{z,\delta} \triangleq \mathbb{1}\left(z_{\mathrm{EW}} > z_\delta \hat{\sigma}_{\varsigma,z}\right).$$

A general discussion regarding $\hat{\sigma}_{\varsigma,z}^2$ is deferred to Section 3.4. When ε is not Gaussian, we only consider the setting where the number of random effects increases to infinity since the analysis of z_{EW} relies of a central limit theorem for quadratic forms.

As mentioned in Section 2.1, under appropriate conditions, we may also use the lasso instead of exponential weighting. For a suitable choice of $\lambda > 0$, define the lasso estimator of β as

$$\hat{\beta}_{\mathrm{LA}} \triangleq \arg\min_{\beta\in\mathbb{R}^p} \|P_W^\perp(Y - X\beta)\|_2^2 + \lambda\|\beta\|_1.$$

Then, the corresponding F-statistic is

$$F_{\mathrm{LA}} \triangleq \frac{\|P_{Z\ominus W}(Y - X\hat{\beta}_{\mathrm{LA}})\|_2^2 / r_{Z\ominus W}}{\|P_{(Z,W)}^\perp(Y - X\hat{\beta}_{\mathrm{LA}})\|_2^2 / r_{(Z,W)^\perp}}.$$

2.3. Assumptions

In this section, we make the following assumptions.

(A1) The mean vector $\mu = \mu_n$ has squared norm, $\|\mu_n\|_2^2/n$, that is, bounded.

(A2) The vector ε is sub-Gaussian with parameter K_ε and has independent components.

(A2*) The vector $\varepsilon \sim \mathcal{N}_n\left(0_n, \sigma_\varepsilon^2 I_n\right)$.

(A3) The random effects ν are sub-Gaussian with parameter K_ν.

(A3*) The random effects ν satisfy $Z\nu \sim \mathcal{N}_n\left(0_n, Z\Psi Z^\mathsf{T}\right)$.

(A4) The matrix Z satisfies $\lambda_{\max}(ZZ^\mathsf{T})$ being bounded and (Z, W) is independent of X.

(A5) The mean vector $\mu = \mu_n$ is weakly sparse relative to X with sparsity s_n, with weak sparsity defined in Definition 1. Furthermore, the statistician chooses a sequence of sparsities u_n such that $u_n \geq s_n$ for n sufficiently large and $u_n = o\left(n^\tau/\log(p)\right)$ for some $\tau \in [1/2, 1]$. Moreover, the number of observations in the reduced models, $r_{Z\ominus W}$ and $r_{(Z,W)^\perp}$, satisfy $r_{Z\ominus W} \asymp r_{(Z,W)^\perp} \asymp n$.

(A6) The mean vector $\mu = \mu_n$ satisfies $\mu = X\beta$ with $\|\beta\|_0 = s_n$. Furthermore, the statistician chooses a sequence of sparsities u_n such that $u_n \geq s_n$ for n sufficiently large and $u_n = o\left(n^\tau/\log(p)\right)$ for some $\tau \in [1/2, 1]$. Moreover, the rows of X are independent and identically distributed $\mathcal{N}_p(0_p, \Sigma_X)$ with $\max(\mathrm{diag}(\Sigma_X)) = \mathcal{O}(1)$ and $r_{(Z,W)^\perp} \asymp n$.

(A7) The vector $\varepsilon = \varepsilon_n$ satisfies

$$\inf_n \left\{\left(\min_{i=1,\ldots,n} \mathrm{Var}(\varepsilon_{n,i})\right) \wedge \left(\min_{i=1,\ldots,n} \mathrm{Var}(\varepsilon_{n,i}^2)\right)\right\} > 0,$$
$$\lim_{x\to\infty} \sup_n \left\{\left(\max_{i=1,\ldots,n} \mathbb{E}\left(\varepsilon_{n,i}^2 : |\varepsilon_{n,i}| > x\right)\right) \vee \left(\max_{i=1,\ldots,n} \mathbb{E}\left(\varepsilon_{n,i}^4 : |\varepsilon_{n,i}| > x\right)\right)\right\} = 0.$$

Remark 1. Assumptions (A1) and (A2) are standard assumptions in high-dimensional linear models. Calling ε the noise and $Z\nu + \varepsilon$ the random component, assumption (A1) is a scaling assumption to ensure the ratio of the fixed components to the random components remains bounded asymptotically and is analogous to the assumptions of Bradic, Claeskens, and Gueuning (2017), who assume that the population covariance matrix of X has bounded maximal eigenvalue and $\|\beta\|_2 = \mathcal{O}(1)$. Next, (A2) is used for consistency of the prediction procedure under the null hypothesis and assumption (A3) allows for concentration of the prediction procedure under the alternative hypothesis. Both assumptions are used by various authors, such as Bradic, Claeskens, and Gueuning (2017) and Cai and Guo (2017). We note that (A2) and (A3) are implied by (A2*) and (A3*), respectively, but the additional Gaussian distribution assumption allows us to relate our methodology to the vast literature on low-dimensional Gaussian mixed models.

Next, the first part of assumption (A4), like (A1), ensures that the ratio of the fixed components to the random components of the variance noise ratio remains bounded under the alternative hypothesis. To elucidate this point, consider the Gaussian setting with $\varepsilon \sim \mathcal{N}_n(0_n, \sigma_\varepsilon^2 I_n)$ and $\nu \sim \mathcal{N}_q(0_q, \sigma_\nu^2 I_q)$. Then, $Z\nu + \varepsilon \sim \mathcal{N}_n(0_n, \sigma_\nu^2 ZZ^\mathsf{T} + \sigma_\varepsilon^2 I_n)$. Since $\lambda_{\max}(ZZ^\mathsf{T}) \geq \max(\mathrm{diag}(ZZ^\mathsf{T}))$, assumption (A4) bounds the variance of the noise. Moreover, (A3) and (A4) together imply that $Z\nu$ is sub-Gaussian with parameter $K_\nu \lambda_{\max}(ZZ^\mathsf{T})$. This requirement is similar to Condition 1 of Bradic, Claeskens, and Gueuning (2017) and Condition 3.1 of Cai and Guo (2017). The assumption that (Z, W) is independent of X is common in the literature (see the discussion before Condition 3.2 of Cai and Guo 2017).

The following two assumptions, (A5) and (A6), are about the sparsity of the fixed effects. The two assumptions consider different asymptotic regimes regarding the random effects; (A5) assumes that the number of random effects increases to infinity while (A6) allows for the number of random effects to stay bounded. The first part of both (A5) and (A6) is a sparsity assumption commonly found in the high-dimensional linear models literature, which is discussed further in Remark 2 below. Note that since the selected sequence of sparsities u_n satisfies $u_n = o(n^\tau/\log(p))$, then the true sequence of sparsities s_n also satisfies the same requirement.

The second half of (A5) is an assumption on the component of the design for the random effects, requiring the number of realizations of the random effects to increase to infinity. The requirement that $r_{Z\ominus W} \asymp r_{(Z,W)^\perp} \asymp n$ is for convenience and can be weakened to only $\min(r_{Z\ominus W}, r_{(Z,W)^\perp}) \to \infty$ if the sparsity requirement is accordingly relaxed to $u_n = o\left(\min(r_{Z\ominus W}, r_{(Z,W)^\perp})^\tau/\log(p)\right)$. The second half of (A6) is a technical requirement to ensure consistency of exponential aggregation for out-of-sample predictions. Since the number of random effects remains bounded, the regression of $P_{Z\ominus W}Y$

on $P_{Z\ominus W}X$ does not necessarily yield a consistent estimator of $P_{Z\ominus W}\mu$ for arbitrary designs. With the Gaussian assumption, we may estimate β by regressing $P^{\perp}_{(Z,W)}Y$ on $P^{\perp}_{(Z,W)}X$ and obtain a consistent estimator of $P_{Z\ominus W}\mu$. Again, the requirement that $r_{(Z,W)^{\perp}} \asymp n$ can be weakened to $r_{(Z,W)^{\perp}} \to \infty$ by adjusting the sparsity requirement to $u_n = o(r_{(Z,W)^{\perp}}{}^{\tau}/\log(p))$.

Assumption (A7) is a mild assumption on the distribution of ε to ensure a central limit theorem. For example, (A7) is satisfied by the Gaussian distribution. Note that no assumption is necessary on γ as the nuisance parameters are projected out in the first stage.

Example 1 (Balanced one-way ANOVA). As an example of a design satisfying the above assumptions on Z, consider a balanced one-way ANOVA design, with q subjects, m observations per subject, and $n = mq$ total observations. In this setting, there are no nuisance random effects, so $d = 0$. Assume further that the number of observations per subject remains bounded (i.e., $m = \mathcal{O}(1)$), which is commonly satisfied in practice. Then, the matrix Z may be represented by $Z = I_q \otimes 1_m$. It is immediate that $r_{Z\ominus W} = q$ and $r_{(Z,W)^{\perp}} = (m-1)q$, implying that the second half of (A5) is satisfied. Finally, assumption (A4) is satisfied since $\lambda_{\max}(ZZ^{\mathsf{T}}) = \lambda_{\max}(mI_q) = m$.

2.4. Main Results

Since F_{EW} is motivated by the classical F-statistic F_{ld}, the following theorem shows that, up to a small bias term depending on the sparsity, the two statistics are asymptotically equivalent.

Theorem 1. Consider the model given in Equation (1) and the hypotheses testing problem from Equation (2). Assume (A1), (A2*), (A3*), (A4), and (A5). If $\alpha \geq 4\left(\sigma_{\varepsilon}^2 + \lambda_{\max}\left(Z\Psi Z^{\mathsf{T}}\right)\right)$, then

$$F_{\mathrm{EW}} = F_{\mathrm{ld}} + o_{\mathbb{P}}(n^{\tau-1}).$$

As mentioned in Section 2.1, under the null hypothesis, the statistic $F_{\mathrm{ld}} \sim F_{r_{Z\ominus(X_S,W)}, r_{(X_S,Z,W)^{\perp}}}$. However, since the weakly sparse set S is unknown, the value of $r_{Z\ominus(X_S,W)}$ and $r_{(X_S,Z,W)^{\perp}}$ cannot be determined in practice. From assumption (A5), as $s = o(n^{\tau}/\log(p))$, then $F_{r_{Z\ominus(X_S,W)}, r_{(X_S,Z,W)^{\perp}}} = F_{r_{Z\ominus W}, r_{(Z,W)^{\perp}}} + o_{\mathbb{P}}(1)$. Thus, the statistic F_{EW} can also be compared to the reference distribution $F_{r_{Z\ominus W}, r_{(Z,W)^{\perp}}}$.

Despite being asymptotically equivalent to the Wald F-test, F_{EW} has an additional bias term of $o_{\mathbb{P}}(n^{\tau-1})$, which impacts the power of the testing procedure. This leads us to consider the following hypotheses testing problem; for any $\tau \in [1/2, 1]$, we consider the contiguous hypotheses

$$\begin{aligned} H_0 &: \lambda_{\max}(P_{Z\ominus W}Z\Psi Z^{\mathsf{T}}P_{Z\ominus W}) = 0, \\ H_1 &: \lambda_{\max, r_{Z\ominus W}}(P_{Z\ominus W}Z\Psi Z^{\mathsf{T}}P_{Z\ominus W}) = hn^{\tau-1}. \end{aligned} \tag{4}$$

Example 2. Consider the setting of Example 1 with ν corresponding to a single random effect and $\nu \sim \mathcal{N}_q(0_q, \sigma_{\nu}^2 I_q)$. Then, with $\tau = 1/2$, the above hypotheses becomes

$$H_0 : \sigma_{\nu}^2 = 0, \qquad H_1 : m\sigma_{\nu}^2 = hn^{-1/2},$$

which is a standard hypotheses testing problem, such as in the balanced one-way random effects model. In this model, in the low-dimensional setting, the rate of $\sqrt{n}$ is optimal.

Theorem 2. Consider the model given by equation (1) and the hypotheses testing problem from equation (4). Assume further (A1), (A2*), (A3*), (A4), and (A5) for any $\tau \in [1/2, 1]$. Fix a value of $\delta > 0$. Under the alternative hypothesis with $h > 0$ sufficiently large (not depending on n) and $\alpha \geq 4\left(\sigma_{\varepsilon}^2 + \lambda_{\max}\left(P_{Z\ominus W}Z\Psi Z^{\mathsf{T}}P_{Z\ominus W}\right)\right)$, the sum of Type I and Type II errors for the test statistic $\phi_{F,\delta}$ is less than one.

Remark 2. The above theorem implies that F_{EW} can distinguish at the classical parametric $\sqrt{n}$ rate if the model is in the ultra-sparse regime, $s = o(\sqrt{n}/\log(p))$. This sparsity rate is common in high-dimensional inference problems for low-dimensional parameters at the parametric rate; in particular, for high-dimensional linear models, a version of this rate is required (cf. Cai and Guo 2017; Javanmard and Montanari 2018). When the value of $\tau \in (1/2, 1]$, we are limited by the ability to remove the bias from the mean vector; in the setting where $\tau = 1/2$, we are limited by the noise level. This seems to suggest a trade-off between the sparsity and the achievable rate of separation.

This comparison with the linear models literature that the inferential procedure requires an additional factor of $\sqrt{n}$ for sparsity assumption appears to be consistent with the recent results by Li, Cai, and Li (2021). In particular, their proposed estimator for the variance components requires a consistent estimator of β. They show in Theorem 3.1 that the minimax rate for estimating β is $s\log(p/s^2)/\mathrm{tr}(\Sigma_a^{-1})$, where $\Sigma_a \in \mathbb{R}^{n\times n}$ is a proxy for the true covariance matrix of Y. Thus, this suggests that $\mathrm{tr}(\Sigma_a^{-1}) \asymp n$, and they require $s\log(p)/n \to 0$ to consistently estimate the variance components.

Remark 3. Compared to the recent work of Li, Cai, and Li (2021), who only suggest an asymptotic distribution for their variance components estimators, Theorem 1 also demonstrates that F_{EW} enjoys certain optimality properties. In addition to providing a distribution under the null hypothesis, Theorem 1 also demonstrates under a sparsity assumption, F_{EW} is asymptotically equivalent to the classical Wald F-test, which is known to enjoy certain optimality properties, such as uniformly most powerful unbiased and uniformly most powerful invariant unbiased in certain ANOVA models (cf. Mathew and Sinha (1988)). In addition, Lu and Zhang (2010) showed that the Wald F-test and likelihood ratio tests are equivalent for balanced one-way ANOVA models while Qeadan and Christensen (2020) showed that the Wald F-test renders the likelihood ratio test inadmissible in generalized split plot designs. Moreover, unlike Li, Cai, and Li (2021), who assume a compatibility condition, our procedure imposes no such requirement on the design matrix X.

We now turn our attention to the setting of sub-Gaussian errors. When $\tau > 1/2$, z_{EW} no longer has an asymptotic Gaussian distribution at the $\sqrt{n}$ rate since the variance dominates the signal. Therefore, in this setting, we only consider hypotheses testing problems as given in Equation (4) with $\tau = 1/2$.

Theorem 3. Consider the model given by equation (1) and the hypotheses testing problem from equation (4). Assume further (A1), (A2), (A3), (A5) for $\tau \leq 1/2$, and (A7). Under the null hypothesis, if $\alpha \geq 4K_\varepsilon$, then

$$\sqrt{n} z_{\mathrm{EW}} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma_{\varsigma,z}^2\right).$$

Remark 4. Compared to Theorem 1, Theorem 3 trades the Gaussian assumption for a sub-Gaussian assumption under a slightly stronger sparsity assumption in order to obtain an asymptotic distribution. From Theorem 2, F_{EW} exhibits a continuous tradeoff between sparsity and power, which does not hold for z_{EW}. This is a consequence of using a central limit theorem for z_{EW}, which requires scaling by $\sqrt{n}$. This implies that the bias should be $o(\sqrt{n})$ and the signal from the alternative should be $\Omega(n^{-1/2})$.

Finally, we end this section by considering the setting where the number of random effects remains bounded.

Theorem 4. Consider the model given in Equation (1) and the hypotheses testing problem from equation (2). Assume (A1), (A2*), (A3*), (A4), and (A6). If $\alpha > 4K_{Z\nu+\varepsilon}$ and $\tilde{\alpha} > 16 \max(\mathrm{diag}(\Sigma_X), \sigma_\varepsilon^2)$, then

$$\tilde{F}_{\mathrm{EW}} = F_{\mathrm{ld}} + o_{\mathbb{P}}(n^{\tau-1}).$$

3. Confidence Intervals for a Single Random Effect

3.1. Model and Motivation

In the previous section, we considered the problem of testing a collection of random effects. However, it is often of interest to construct confidence intervals for the variance of a particular random effect. Suppose that $\Psi = \sigma_\nu^2 I_\nu$. In the low-dimensional setting, there have been many procedures suggested to construct confidence intervals, from likelihood based approaches to F-test inversions (see, e.g., Jiang (2007) for a nonexhaustive list). In this section, we deal with a confidence interval for a single variance component, which can easily be extended using a Bonferroni correction or similar procedures for simultaneous confidence intervals. Alternatively, we may also invert the F-statistic from Section 2 to obtain confidence intervals for parameters of the form $\sigma_\nu^2/\sigma_\varepsilon^2$, with such ratios being first studied by Hartley and Rao (1967).

Our high-dimensional approach is inspired by F-test inversion. However, instead of using the ratio, we again use the difference. Define

$$Q \triangleq \begin{pmatrix} P_{Z\ominus W} - r_{Z\ominus W} r_{(Z,W)^\perp}{}^{-1} P_{(Z,W)}^\perp & P_{Z\ominus W} Z \\ Z^\mathsf{T} P_{Z\ominus W} & Z^\mathsf{T} P_{Z\ominus W} Z \end{pmatrix},$$
$$\xi \triangleq \begin{pmatrix} \varepsilon \\ \nu \end{pmatrix}.$$

Then, expanding the statistic z_{EW} from Section 2, we have that

$$\begin{aligned} z_{\mathrm{EW}} &= \|P_{Z\ominus W}(Y - X\hat{\beta}_{\mathrm{EW}})\|_2^2 - r_{Z\ominus W} r_{(Z,W)^\perp}{}^{-1} \\ &\quad \times \|P_{(Z,W)}^\perp (Y - X\hat{\beta}_{\mathrm{EW}})\|_2^2 \\ &= \|P_{Z\ominus W}(Z\nu + \varepsilon)\|_2^2 - r_{Z\ominus W} r_{(Z,W)^\perp}{}^{-1} \\ &\quad \times \|P_{(Z,W)}^\perp \varepsilon\|_2^2 + o_{\mathbb{P}}(n^\tau) \\ &= \|P_{Z\ominus W}(Z\nu + \varepsilon) - r_{Z\ominus W}{}^{1/2} r_{(Z,W)^\perp}{}^{-1/2} P_{(Z,W)}^\perp \varepsilon\|_2^2 \\ &\quad + o_{\mathbb{P}}(n^\tau) \\ &= \xi^\mathsf{T} Q \xi + o_{\mathbb{P}}(n^\tau), \end{aligned}$$

where the second equality follows from Lemma S1 in the supplementary material. A direct calculation shows that $\mathbb{E}\xi^\mathsf{T} Q \xi = \sigma_\nu^2 \mathrm{tr}(Z^\mathsf{T} P_{Z\ominus W} Z)$. Then, with proper centering and scaling, we may apply a central limit theorem for quadratic forms under a mild condition on the matrix Q.

3.2. Estimator

To estimate σ_ν^2, we consider $\hat{\sigma}_\nu^2$ defined by

$$\begin{aligned} \hat{\sigma}_\nu^2 \triangleq [\mathrm{tr}(Z^\mathsf{T} P_{Z\ominus W} Z)]^{-1} \Big(&\|P_{Z\ominus W}(Y - X\hat{\beta}_{\mathrm{EW}})\|_2^2 \\ &- r_{Z\ominus W} r_{(Z,W)^\perp}{}^{-1} \|P_{(Z,W)}^\perp (Y - X\hat{\beta}_{\mathrm{EW}})\|_2^2 \Big). \end{aligned}$$

By a direct calculation, it can be shown that

$$\begin{aligned} \sigma_\varsigma^2 \triangleq \mathrm{Var}(\xi^\mathsf{T} Q \xi) = \kappa_\varepsilon \sum_{i=1}^n Q_{i,i}^2 + \kappa_\nu \sum_{i=n+1}^{n+q} Q_{i,i}^2 \\ + 2 \sum_{i \neq j} Q_{i,j}^2 (\sigma_\varepsilon^2 \mathbb{1}(1 \leq i \leq n) \\ + \sigma_\nu^2 \mathbb{1}(n+1 \leq i \leq n+q)) \\ \times (\sigma_\varepsilon^2 \mathbb{1}(1 \leq j \leq n) \\ + \sigma_\nu^2 \mathbb{1}(n+1 \leq j \leq n+q)). \end{aligned}$$

From the above, we see that the asymptotic distribution of $\hat{\sigma}_\nu^2$ depends on the second and fourth moments of ν and ε. To estimate the second moment of ε, we consider the estimator

$$\hat{\sigma}_\varepsilon^2 \triangleq r_{(Z,W)^\perp}{}^{-1} \|P_{(Z,W)}^\perp (Y - X\hat{\beta}_{\mathrm{EW}})\|_2^2.$$

The problem of estimation of fourth moments requires some technical assumptions on the design, even in the low-dimensional setting. For simplicity, we only consider the setting of Gaussian mixed models and the balanced one-way ANOVA design, but we note that the arguments may be extended under suitable regularity on the design matrices Z and W. In the setting, Gaussian mixed models, the fourth moment is entirely determined by the second moment. For the setting of the balanced one-way ANOVA design with m observations per subject, we consider the estimator

$$\begin{aligned} \hat{\omega}_\varepsilon &\triangleq q^{-1} m^2 \|P_{(Z,W)}^\perp (Y - X\hat{\beta}_{\mathrm{EW}})\|_4^4 - 3(m-1)\hat{\sigma}_\varepsilon^4, \\ \hat{\omega}_\nu &\triangleq (mq)^{-1} \|P_{Z\ominus W}(Y - X\hat{\beta}_{\mathrm{EW}})\|_4^4 - 6m^{-1}\hat{\sigma}_\varepsilon^2 \hat{\sigma}_\nu^2 \\ &\quad - m^{-3}\hat{\omega}_\varepsilon - 3m^{-3}(m-1)\hat{\sigma}_\varepsilon^4, \\ \hat{\kappa}_\varepsilon &\triangleq \hat{\omega}_\varepsilon - \hat{\sigma}_\varepsilon^4, \qquad \hat{\kappa}_\nu \triangleq \hat{\omega}_\nu - \hat{\sigma}_\nu^4. \end{aligned}$$

In both settings, we obtain a plug-in estimator $\hat{\sigma}_\varsigma^2$ of σ_ς^2. By setting $\hat{\kappa}_\nu = 0$ and $\hat{\sigma}_\nu^2 = 0$, we obtain an estimator $\hat{\sigma}_{\varsigma,z}^2$ of $\sigma_{\varsigma,z}^2$ for Section 2.

Remark 5. The statistic $\hat{\sigma}_\nu^2$ is related to the classical analysis of variance method for estimating random effects. Consider the setting of a balanced one-way ANOVA model from Example 1 with $\mu = 0_n$. Let

$$\begin{aligned} MS_{\mathrm{Treatments}} &= q^{-1} \|P_Z Y\|_2^2 = q^{-1} \|P_Z (Z\nu + \varepsilon)\|_2^2, \\ MS_{\mathrm{Error}} &= (n-q)^{-1} \|P_Z^\perp Y\|_2^2 = (n-q)^{-1} \|P_Z^\perp \varepsilon\|_2^2. \end{aligned}$$

Then, the analysis of variance estimate is given by

$$\hat{\sigma}_{\nu,\mathrm{AOV}}^2 \triangleq m^{-1}(MS_{\mathrm{Treatments}} - MS_{\mathrm{Error}}).$$

Now, note that $r_{Z\ominus W} = q$, $r_{(Z,W)^\perp} = (m-1)q$, $n = mq$, and $\mathrm{tr}(Z^\mathsf{T} P_Z Z) = \mathrm{tr}(Z^\mathsf{T} Z) = mq$. From the calculations in Section 3.1, we have that

$$\hat{\sigma}_\nu^2 = (mq)^{-1}\left(\|P_Z(Z\nu+\varepsilon)\|_2^2 - (m-1)^{-1}\|P_Z^\perp \varepsilon\|_2^2 + o_\mathbb{P}(q^\tau)\right) = \hat{\sigma}_{\nu,\mathrm{AOV}}^2 + o_\mathbb{P}(1).$$

Thus, the two statistics are asymptotically equivalent in the balanced one-way ANOVA setting.

3.3. Assumptions

In addition to the assumptions from Section 2, we need additional assumptions on the matrix Q and on the distribution of the random effects ν.

(B1) The matrix Q satisfies

$$\frac{\lambda_{\max}(Q^2)}{\mathrm{tr}(Q^2)} \to 0.$$

(B2) The vector $\nu = \nu_n$ satisfies

$$\begin{aligned}&\inf_n \left\{\left(\min_{i=1,\dots,q_n} \mathrm{Var}(\nu_{n,i})\right) \wedge \left(\min_{i=1,\dots,q_n} \mathrm{Var}(\nu_{n,i}^2)\right)\right\} > 0,\\ &\lim_{x\to\infty} \sup_n \left\{\left(\max_{i=1,\dots,q_n} \mathbb{E}\left(\nu_{n,i}^2 : |\nu_{n,i}| > x\right)\right) \vee \left(\max_{i=1,\dots,q_n} \mathbb{E}\left(\nu_{n,i}^4 : |\nu_{n,i}| > x\right)\right)\right\} = 0.\end{aligned}$$

Remark 6. Assumptions (B1) and (B2), along with (A7), are used for a central limit theorem for quadratic forms. For a thorough discussion on these assumptions, we refer the interested reader to Section 5 of Jiang (1996). As a consequence of using a central limit theorem, we require that the number of random effects increases to infinity. Thus, we only consider the sparsity assumption (A5) in this section.

Example 3 (Balanced one-way ANOVA (ctd.)). Continuing with Example 1, we note that $ZZ^\mathsf{T} = mP_{Z\ominus W}$. Also, recall that $r_{Z\ominus W} = q$ and $r_{(Z,W)^\perp} = (m-1)q$. Then,

$$Q^2 = \begin{pmatrix} (m+1)P_{Z\ominus W} + (m-1)^{-2}P_{(Z,W)}^\perp & (m+1)Z \\ (m+1)Z^\mathsf{T} & (m^2+m)I_q \end{pmatrix}.$$

A direct calculation shows that $\lambda_{\max}(Q^2) = (m+1)^2$ and $\mathrm{tr}(Q^2) = (m+1)q + (m-1)^{-1}q + (m^2+m)q$, which satisfies assumption (B1).

3.4. Main Results

We start by stating the asymptotic distribution of $\hat{\sigma}_\nu^2$.

Theorem 5. Consider the model in Equation (1). Assume (A1), (A2), (A3) with $\Psi = \sigma_\nu^2 I_q$, (A4), (A5) with $\tau = 1/2$, (A7), (B1), and (B2). If $\alpha > 4(K_\nu \lambda_{\max}(ZZ^\mathsf{T}) + K_\varepsilon)$, then

$$\sigma_\varsigma^{-1}[\mathrm{tr}(Z^\mathsf{T} P_{Z\ominus W} Z)](\hat{\sigma}_\nu^2 - \sigma_\nu^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Next, we consider the following lemma, which shows that $\hat{\kappa}_\varepsilon$ and $\hat{\kappa}_\nu$ are consistent estimators of κ_ε and κ_ν.

Proposition 1. Consider the balanced one-way ANOVA from Example 1. Under the assumptions of Theorem 5,

$$\hat{\kappa}_\varepsilon \xrightarrow{\mathbb{P}} \kappa_\varepsilon, \qquad \hat{\kappa}_\nu \xrightarrow{\mathbb{P}} \kappa_\nu.$$

Thus, the preceding two results allow us to construct confidence intervals in the Gaussian mixed model and the balanced one-way ANOVA setting. Let $\hat{\sigma}_\varsigma^2$ be a consistent estimator for σ_ς^2. Then, an asymptotic $(1-\delta)$ confidence interval for σ_ν^2 may be given by

$$(\hat{\sigma}_\nu^2 - z_\delta \hat{\sigma}_\varsigma [\mathrm{tr}(Z^\mathsf{T} P_{Z\ominus W} Z)]^{-1}, \hat{\sigma}_\nu^2 + z_\delta \hat{\sigma}_\varsigma [\mathrm{tr}(Z^\mathsf{T} P_{Z\ominus W} Z)]^{-1})$$

where z_δ is the δ upper quantile of a standard Gaussian distribution. Since the above interval may be negative, we may truncate negative values to zero.

4. Empirical Bayes in ANOVA Type Models

The motivating example of this problem framework is in terms of the Rasch model, originally proposed by Rasch (1960). The model that we consider is different than the classical Rasch model in that we have Gaussian responses as opposed to binary responses. Our interest in this section is not in testing whether the variance of the random effect is different from zero, but, assuming that it is different from zero, in estimating the individual components of the random effect. We use the term empirical Bayes, or compound decision, in the sense of Efron (2019) and the references therein (specifically Greenshtein and Ritov 2019).

As an example of this model, the data that we consider in Section 6 is from the Trends in Mathematics and Sciences Study (TIMSS), an international study conducted every four years to measure fourth and eighth grade student achievement in mathematics and science. We only consider data from the year 2015. Polities randomly sample a collection of nationally representative schools to take standardized examinations in both mathematics and science, with questions being either multiple choice or constructed response. Then, each student within schools takes only a subset of the questions on the exams but all questions are answered by some students in each school. In addition to recording student responses, the data also contains background covariates for schools. Martin, Mullis, and Hooper (2016) provides a more detailed description of the methods and procedures employed by TIMSS and more general information about TIMSS is available in Mullis, Martin, and Loveless (2016).

For our analysis, we only consider multiple choice questions and analyze on the level of school rather than students. To construct a response variable for school, we compute the proportion of questions answered correctly by students in that school. Note that, unlike the classical Rasch model, we assume a linear model and, for all schools, we have answers for all questions. Thus, by a central limit theorem, our response Y is approximately Gaussian. The fixed effects design X include the background covariates for the school and the random effects design Z is an indicator for the polity, with ν corresponding to the unobserved variability of the polities. In this example, since we have averaged over questions, we do not have any nuisance random effects. The problem that we consider in this section is ranking the polities based on mathematical ability and trying to estimate the average number of questions that any particular polity will answer correctly. That is, we would like to estimate $\mu + Z\nu$ for all polities in our dataset.

4.1. Model and Motivation

The general problem framework that we consider is for K-factor ANOVA models. However, we derive the results in the setting when $K = 2$. That is, we consider the model

$$Y = \mu + Z\nu + W\gamma + \varepsilon.$$

We do not assume that the design is fully crossed in the random effects. The goal in the problem is to estimate a subset of the mean vector, $\eta \triangleq \mu + Z\nu$, since we view the random effects W as nuisance. However, as the sample size increases, the number of observations per group stays bounded. In the context of the motivating data example, each school still only answers a finite number of questions as we increase the sample size. A standard approach in the low-dimensional setting would be to use an empirical Bayes estimator by placing a Gaussian prior on both ν and γ (see, e.g., Brown, Mukherjee, and Weinstein 2018), which transforms the problem into a standard high-dimensional linear mixed model. Therefore, we use a $\mathcal{N}_\nu\left(0_\nu, \sigma_\nu^2 I_\nu\right)$ and $\mathcal{N}_r\left(0_r, \sigma_\gamma^2 I_r\right)$ prior on ν and γ, respectively.

Since we need to estimate both σ_ν^2 and σ_γ^2 for the prior, our estimator for σ_γ^2 is analogous to $\hat{\sigma}_\nu^2$ from Section 3. To this end, we need an additional matrix $P_{W\ominus Z}$ such that

$$P_{W\ominus Z}Y = P_{W\ominus Z}X\beta + P_{W\ominus Z}W\gamma + P_{W\ominus Z}\varepsilon.$$

4.2. Estimator

Since we are also interested in estimating $P_{W\ominus Z}X\beta$, we define $\tilde{\beta}_{\mathrm{EW}}$ to be the exponentially weighted estimator using the covariates X, as opposed to using the covariates $P_W^\perp X$ for $\hat{\beta}_{\mathrm{EW}}$. Then, analogous to Section 2.2 let $\tilde{\beta}_m$ denote the least-squares estimator of β using the model $m \in \mathcal{M}_u$ with covariates X_m and $K_{Z\nu+W\gamma+\varepsilon}$ be the sub-Gaussian parameter for $Z\nu+W\gamma+\varepsilon$. For $\tilde{\alpha} > 4K_{Z\nu+W\gamma+\varepsilon}$, defining the exponential weights as

$$\tilde{w} \triangleq \frac{\exp\left(-\frac{1}{\tilde{\alpha}}\|Y - X\tilde{\beta}_m\|_2^2\right)}{\sum_{k\in\mathcal{M}_u}\exp\left(-\frac{1}{\tilde{\alpha}}\|Y - X\tilde{\beta}_k\|_2^2\right)},$$

we have

$$\tilde{\beta}_{\mathrm{EW}} \triangleq \sum_{m\in\mathcal{M}_u}\tilde{w}_m\tilde{\beta}_m.$$

For convenience, we write $\tilde{\mu}_{\mathrm{EW}} \triangleq X\tilde{\beta}_{\mathrm{EW}}$. Now, the estimators for the variance are given by

$$\begin{aligned}
\tilde{\sigma}_\nu^2 &\triangleq [\mathrm{tr}(Z^\mathsf{T}P_{Z\ominus W}Z)]^{-1}\Big(\|P_{Z\ominus W}(Y - \tilde{\mu}_{\mathrm{EW}})\|_2^2 \\
&\quad - r_{Z\ominus W}{r_{(Z,W)^\perp}}^{-1}\|P_{(Z,W)}^\perp(Y - \tilde{\mu}_{\mathrm{EW}})\|_2^2\Big), \\
\tilde{\sigma}_\gamma^2 &\triangleq [\mathrm{tr}(W^\mathsf{T}P_{W\ominus Z}W)]^{-1}\Big(\|P_{W\ominus Z}(Y - \tilde{\mu}_{\mathrm{EW}})\|_2^2 \\
&\quad - r_{W\ominus Z}{r_{(Z,W)^\perp}}^{-1}\|P_{(Z,W)}^\perp(Y - \tilde{\mu}_{\mathrm{EW}})\|_2^2\Big), \\
\tilde{\sigma}_\varepsilon^2 &\triangleq {r_{(Z,W)^\perp}}^{-1}\|P_{(Z,W)}^\perp(Y - \tilde{\mu}_{\mathrm{EW}})\|_2^2.
\end{aligned}$$

As we do not require an asymptotic distribution for $\tilde{\sigma}_\nu^2$ and $\tilde{\sigma}_\gamma^2$, under weaker assumptions that Theorem 5, we have that $\tilde{\sigma}_\nu^2$ and $\tilde{\sigma}_\gamma^2$ are consistent estimators of σ_ν^2 and σ_γ^2, respectively. This suggests the the following empirical Bayes estimator for η,

$$\tilde{\eta}_{\mathrm{EW}} \triangleq \tilde{\mu}_{\mathrm{EW}} + \tilde{\sigma}_\nu^2 ZZ^\mathsf{T}\left(\tilde{\sigma}_\nu^2 ZZ^\mathsf{T} + \tilde{\sigma}_\gamma^2 WW^\mathsf{T} + \tilde{\sigma}_\varepsilon^2 I_n\right)^{-1}(Y - \tilde{\mu}_{\mathrm{EW}}).$$

To compare our estimator, we consider an oracle that has access to μ, σ_ν^2, σ_γ^2, and σ_ε^2. Then, this oracle uses the Bayes estimator for η (see Lemma 6), given by

$$\tilde{\eta}_{\mathrm{oracle}} \triangleq \mu + \sigma_\nu^2 ZZ^\mathsf{T}\left(\sigma_\nu^2 ZZ^\mathsf{T} + \sigma_\gamma^2 WW^\mathsf{T} + \sigma_\varepsilon^2 I_n\right)^{-1}(Y - \mu).$$

4.3. Assumptions

As previously mentioned, we do not need to establish the asymptotic distribution of $\hat{\sigma}_\nu^2$, rather we only need the estimator to be consistent. Accordingly, we may weaken our assumptions to the following

(C1) The designs Z and W satisfy $\mathrm{tr}(Z^\mathsf{T}P_{Z\ominus W}Z) \asymp \mathrm{tr}(W^\mathsf{T}P_{W\ominus Z}W) \asymp n$.
(C2) The matrix W satisfies $\lambda_{\max}(WW^\mathsf{T})$ being bounded.

Remark 7. Assumption (C1) ensures that the component of the design for the random effects is sufficiently well balanced. This assumption in the presence of (A4) implies the second half of (A5). Note that $\mathrm{tr}(Z^\mathsf{T}P_{Z\ominus W}Z) \le \lambda_{\max}(ZZ^\mathsf{T})\mathrm{tr}(P_{Z\ominus W}) = \lambda_{\max}(ZZ^\mathsf{T})r_{Z\ominus W}$. Since $r_{Z\ominus W} \le n$, $\mathrm{tr}(Z^\mathsf{T}P_{Z\ominus W}Z) \asymp n$ and $\lambda_{\max}(ZZ^\mathsf{T})$ being bounded imply that $r_{Z\ominus W} \asymp n$.

The other assumption (C2) is analogous to (A4).

4.4. Main Results

We start this section by noting that $\hat{\sigma}_\nu^2$, $\hat{\sigma}_\gamma^2$, and $\hat{\sigma}_\varepsilon^2$ are all consistent estimators under a weaker sparsity assumption than in Section 3. Since we no longer require an asymptotic distribution for the variance estimates, we only need the prediction rate to ensure consistency, which is the content of the ensuing proposition.

Proposition 2. Consider the model given in Equation (1). Assume (A1), (A2*), (A3*) with $\Psi = \sigma_\nu^2 I_\nu$, (A4), (A5) with $\tau = 1$, (C1), and (C2). If $\alpha > 4(\sigma_\nu^2\lambda_{\max}(ZZ^\mathsf{T}) + \sigma_\gamma^2\lambda_{\max}(WW^\mathsf{T}) + \sigma_\varepsilon^2)$, then

$$\tilde{\sigma}_\nu^2 \xrightarrow{\mathbb{P}} \sigma_\nu^2, \qquad \tilde{\sigma}_\gamma^2 \xrightarrow{\mathbb{P}} \sigma_\gamma^2, \qquad \tilde{\sigma}_\varepsilon^2 \xrightarrow{\mathbb{P}} \sigma_\varepsilon^2.$$

The following is a standard lemma regarding the empirical Bayes estimators in this problem setup, which we prove for the sake of completeness.

Lemma 6. For a fixed vector $\mu \in \mathbb{R}^n$ and fixed values $\sigma_\nu^2 > 0$, $\sigma_\gamma^2 > 0$, and $\sigma_\varepsilon^2 > 0$, the Bayes estimator of η is given by

$$\mathbb{E}(\eta|Y) = \mu + \sigma_\nu^2 ZZ^\mathsf{T}\left(\sigma_\nu^2 ZZ^\mathsf{T} + \sigma_\gamma^2 WW^\mathsf{T} + \sigma_\varepsilon^2 I_n\right)^{-1}(Y - \mu).$$

We conclude this section with the main result regarding $\tilde{\eta}_{\mathrm{EW}}$; the empirical Bayes estimator performs nearly as well as the oracle Bayes estimator $\tilde{\eta}_{\mathrm{oracle}}$ asymptotically.

Theorem 7. Consider the model given in Equation (1). Under the assumptions of Proposition 2,

$$n^{-1}\left(\|\tilde{\eta}_{\text{EW}}-\eta\|^2-\|\tilde{\eta}_{\text{oracle}}-\eta\|^2\right)=o_{\mathbb{P}}(1).$$

5. Simulations

5.1. Methods and Models

We consider the linear mixed model given by

$$Y=X\beta+Z\nu+W\gamma+\varepsilon,$$

with $n=1000$, $p=2000$, and $q=200$. The parameters that we vary throughout the experiment are the sparsity s, the distribution of X, ν, γ, and ε, the value of σ_ν^2, and the number of nuisance random effects d. For each parameter setting, the results are averaged over 100 replications.

For the sparsity, we set $s\in\{3,15\}$. Each row of X is independent and identically distributed $\mathcal{N}_p\left(0_p,\Sigma\right)$ with

$$\Sigma_{i,j}=\begin{cases}1 & \text{if } i=j,\\ \rho & \text{if } i\neq j,\end{cases}$$

for $\rho\in\{0,0.8\}$. Then, β is chosen such that the signal strength, $\beta^{\mathsf{T}}\Sigma\beta$, is four times the noise level with $\sigma_\varepsilon^2=1$. This is accomplished by first generating s uniform random variables in $[-1,1]$ and then rescaling to the desired level.

For the random effects, we either generate them from a Gaussian distribution or a double exponential distribution, which we denote by "z" and "e," respectively. For the variances, we let $\sigma_\nu^2\in\{0,1\}$ while $\sigma_\gamma^2=1$.

Finally, for the component of the design corresponding to the random effects, we let $d\in\{0,200\}$. When $d=0$, the design is a balanced one-way ANOVA design with $m=5$. When $d=200$, we generate from a two-way crossed design and down sample to have n observations. We only consider the sub-Gaussian procedures when $d=0$.

All of our simulations are conducted in R. For each of our three problems, we compare the exponential weighting estimator, denoted by "EW" with an oracle low-dimensional estimator as well as a low-dimensional version of our proposed high-dimensional statistic.

For exponential weighting, we follow Algorithm 1 from Law and Ritov (2021). Regarding the tuning parameters, we perform four fold cross-validation over a grid of values for α and the sparsity.

For the oracle estimators, in the setting of the F-test, we directly apply the classical low-dimensional F-test that has access to the true sparse set S, as given in Equation (2.3) of Jiang (2007). For the confidence intervals, we fit the linear mixed models with the true sparse set S using `lmer` and applying the `confint` function. Finally, in the setting of estimation, we directly compute the oracle Bayes estimator $\tilde{\eta}_{\text{oracle}}$ described in Section 4. Collectively, these low-dimensional estimators are denoted by "LD."

In addition to comparing with the low-dimensional estimators, we also construct low-dimensional versions of our proposed high-dimensional statistics. To do so, we use the exact same statistic as in the high-dimensional setting but replace exponential weighting with least-squares using the sparse set S. We make this comparison since all of our proposed statistics rely on two layers of asymptotics:

1. In the prediction of the mean vector via exponential weighting.
2. In the convergence once the residuals are obtained.

To differentiate between these two, we introduce an intermediate statistic that relies on least-squares, which we think of as low-dimensional versions of our statistics. For example, letting $\hat{\beta}_S$ be the least-squares estimator of β using the covariates X_S, we also consider the statistic

$$F_{\text{LS}}\triangleq\frac{\|P_{Z\ominus W}(Y-X\hat{\beta}_S)\|^2/r_{Z\ominus W}}{\|P_{(Z,W)}^{\perp}(Y-X\hat{\beta}_S)\|^2/r_{(Z,W)^\perp}}\sim F_{r_{Z\ominus W},r_{(Z,W)^\perp}}+o_{\mathbb{P}}(1).$$

These estimators are denoted by "LS."

Finally, we also include a version of our statistics using scaled lasso, which we denote by "SL." Then, for "EW," "SL," and "LS," we subscript them by either "G" or "SG" to distinguish between the Gaussian and sub-Gaussian methods.

To compare the procedures, we consider the following metrics

1. Type I/II Error: The percentage of time the procedure produces a Type I or Type II error in hypothesis testing.
2. Average Coverage: The percentage of time the correct hypothesis is selected for F-tests or the percentage of time the true value of σ_ν^2 is in the confidence interval.
3. Average Length: The average length of the confidence interval, taken as the upper endpoint minus the lower endpoint.
4. Average Loss: The average squared Euclidean distance between the estimated vector $\hat{\eta}$ and the true vector η divided by n.

5.2. Results

The results are presented in Tables S1–S3 from the supplementary material. We notice that for hypothesis testing, all the procedures control Type I and Type II error well throughout the settings. For confidence intervals, when $d=0$ and $s=3$, we notice that all of the methods perform well in coverage. However, the length of our procedures appears to be shorter when $\sigma_\nu^2=0$ and longer when $\sigma_\nu^2=1$, whereas the low-dimensional procedure is more uniform across the parameter space. This is not surprising in view of our estimation procedure. From Section 3.2, the asymptotic variance of $\hat{\sigma}_\nu^2$ depends monotonically on the second and fourth moments of ν and ε, which is reflected in the lengths of the resulting intervals.

When $\sigma_\nu^2=0$, the empirical coverage of our confidence intervals are close to the nominal level, even when the distribution of the random effects and errors are double exponential. When $\sigma_\nu^2=1$, the empirical coverage drops to around 80% for the Gaussian procedure and 90% for the sub-Gaussian procedure when the distribution is double exponential, against a nominal coverage of 95%. We note that the double exponential distribution is not a sub-Gaussian distribution, which seems to suggest that the confidence intervals are somewhat robust to slight departures from the distributional assumptions.

Moreover, when increasing the sparsity from $s=3$ to $s=15$, the performance of our confidence intervals decreases slightly

since it is harder to remove the contribution of the fixed effects. Finally, for empirical Bayes estimation, our methods are competitive with the oracle. However, we notice that exponential weighting outperforms scaled lasso when $s = 15$, particularly when $\rho = 0.8$. Since larger values of ρ implies that the columns of X are more correlated, this highlights a salient feature of exponential weighting.

6. Real Data Application

Following in the motivating example of Section 4, we consider the TIMSS dataset, which is freely available at *https://timssandpirls.bc.edu/*. To simplify our analysis, we only consider the mathematics questions. After filtering out for complete cases on background covariates, we are left with 146 questions, $q = 43$ unique polities, $p = 106$ covariates, and 6808 schools. Therefore, we had a total of $n = 6808$ responses after averaging over the students and questions within the schools. Here, there are no nuisance random effects so $d = 0$. Due to averaging over students within schools, we expect the distributions to be approximately Gaussian by a central limit theorem.

To demonstrate our methodology, we use both exponential weighting as well as scaled lasso as our estimation procedure. When applying exponential weighting, we jointly tune the value of u and α using four fold cross-validation. The high-dimensional F-test rejected the null hypothesis that $\sigma_\nu^2 = 0$ and a 95% confidence interval for σ_ν^2 is $(0.0021, 0.0056)$, which suggests that, even controlling for school background characteristics, the polity of the school impacts mathematical ability. For the last part, we define a polity's background characteristics X to be the arithmetic average of all the schools' background characteristics within that polity. Then, applying the empirical Bayes procedure, we rank the polities based on the predicted number of questions they would answer correctly. The top five polities in order from our analysis are South Korea, Singapore, Hong Kong, Chinese Taipei, and Japan. Up to some reordering, our results are mostly consistent with the report of Mullis et al. (2016) based on individual student data, who had the same top five polities. The results using scaled lasso produced the same ranking as exponential weighting and similar conclusions regarding σ_ν^2.

Supplementary Materials

In the supplement, we provide proofs for all of the results along with additional simulation tables.

Funding

Supported in part by NSF Grants DMS-1646108 and DMS-1712962.

References

Bradic, J., Claeskens, G., and Gueuning, T. (2017), "Fixed Effects Testing in High-Dimensional Linear Mixed Models," *Journal of the American Statistical Association*, 115(532), 1835–1850. [1682,1685]

Brown, L. D., Mukherjee, G., and Weinstein, A. (2018), "Empirical Bayes Estimates for a Two-Way Cross-Classified Model," *The Annals of Statistics*, 46, 1693–1720. [1689]

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. [1685,1686]

Chen, F., Li, Z., Shi, L., and Zhu, L. (2015), "Inference for Mixed Models of ANOVA Type with High-Dimensional Data," *Journal of Multivariate Analysis*, 133, 382–401. [1682,1683]

Efron, B. (2019), "Bayes, Oracle Bayes and Empirical Bayes," *Statistical Science*, 34, 177–201. [1688]

Greenshtein, E., and Ritov, Y. (2019), "Comment: Empirical Bayes, Compound Decisions and Exchangeability," *Statistical Science*, 34, 224–228. [1688]

Groll, A., and Tutz, G. (2014), "Variable Selection for Generalized Linear Mixed Models by l1-Penalized Estimation," *Statistics and Computing*, 24, 137–154. [1682]

Hartley, H. O., and Rao, J. N. (1967), "Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model," *Biometrika*, 54, 93–108. [1687]

Javanmard, A., and Montanari, A. (2018), "Debiasing the Lasso: Optimal Sample Size for Gaussian Designs," *The Annals of Statistics*, 46, 2593–2622. [1686]

Jiang, J. (1996), "Reml Estimation: Asymptotic Behavior and Related Topics," *The Annals of Statistics*, 24, 255–286. [1688]

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, New York: Springer Science and Business Media. [1687,1690]

Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H. (2016), "On High-Dimensional Misspecified Mixed Model Analysis in Genome-Wide Association Study," *The Annals of Statistics*, 44, 2127–2160. [1682]

Law, M., and Ritov, Y. (2021), "Inference Without Compatibility: Using Exponential Weighting for Inference on a Parameter of a Linear Model," *Bernoulli*, 27(3), 1467–1495. https://doi.org/10.3150/20-BEJ1280. [1683,1684,1690]

Leung, G., and Barron, A. R. (2006), "Information Theory and Mixing Least-Squares Regressions," *IEEE Transactions on Information Theory*, 52, 3396–3410. [1684]

Li, S., Cai, T. T., and Li, H. (2021), "Inference for High-Dimensional Linear Mixed-Effects Models: A Quasi-Likelihood Approach," *Journal of the American Statistical Association*, 1–33. https://doi.org/10.1080/01621459.2021.1888740. [1682,1686]

Lu, Y., and Zhang, G. (2010), "The Equivalence Between Likelihood Ratio Test and f-test for Testing Variance Component in a Balanced One-Way Random Effects Model," *Journal of Statistical Computation and Simulation*, 80, 443–450. [1686]

Martin, M. O., Mullis, I. V., and Hooper, M. (2016), "Methods and Procedures in Timss 2015," *TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA)*. [1688]

Mathew, T., and Sinha, B. K. (1988), "Optimum Tests in Unbalanced Two-Way Models Without Interaction," *The Annals of Statistics*, 16, 1727–1740. [1686]

Müller, S., Scealy, J. L., and Welsh, A. H. (2013), "Model Selection in Linear Mixed Models," *Statistical Science*, 28, 135–167. [1682]

Mullis, I., Martin, M., Foy, P., and Hooper, M. (2016), "Timss 2015 International Results in Mathematics," Retrieved from Boston College, TIMSS and PIRLS International Study Center. [1691]

Mullis, I. V., Martin, M. O., and Loveless, T. (2016), "20 Years of TIMSS: International Trends in Mathematics and Science Achievement, Curriculum, and Instruction," *TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA)*. [1688]

Qeadan, F., and Christensen, R. (2020), "On the Equivalence Between the lrt and f-test for Testing Variance Components in a Class of Linear Mixed Models," *Metrika: International Journal for Theoretical and Applied Statistics*, 84, 313–338. [1686]

Rasch, G. (1960), "Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests." [1688]

Schelldorfer, J., Bühlmann, 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. [1682]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society* Series B, 58, 267–288. [1684]

Wang, L., Zhou, J., and Qu, A. (2012), "Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis," *Biometrics*, 68, 353–360. [1682]