Confidence Regions Near Singular Information and Boundary Points With Applications to Mixed Models

Karl Oskar Ekvall*
karl.oskar.ekvall@ki.se

Matteo Bottai
matteo.bottai@ki.se

Division of Biostatistics, Institute of Environmental Medicine, Karolinska Institute
Nobels väg 13, 17177 Stockholm, Sweden

March 18, 2021

Abstract

We propose confidence regions with asymptotically correct uniform coverage probability of parameters whose Fisher information matrix can be singular at important points of the parameter set. Our work is motivated by the need for reliable inference on scale parameters close or equal to zero in mixed models, which is obtained as a special case. The confidence regions are constructed by inverting a continuous extension of the score test statistic standardized by expected information, which we show exists at points of singular information under regularity conditions. Similar results have previously only been obtained for scalar parameters, under conditions stronger than ours, and applications to mixed models have not been considered. In simulations our confidence regions have near-nominal coverage with as few as $n = 20$ observations, regardless of how close to the boundary the true parameter is. It is a corollary of our main results that the proposed test statistic has an asymptotic chi-square distribution with degrees of freedom equal to the number of tested parameters, even if they are on the boundary of the parameter set.

*Corresponding author

1 Introduction

In mixed effects models, the importance of a random effect is often assessed by inference on a variance or scale parameter. A parameter near zero typically indicates a weak effect and many tests for whether a variance is equal to zero have been proposed (Stram and Lee, 1994; Lin, 1997; Stern and Welsh, 2000; Hall and Praestgaard, 2001; Verbeke and Molenberghs, 2003; Crainiceanu and Ruppert, 2004; Zhu and Zhang, 2006; Fitzmaurice et al., 2007; Greven et al., 2008; Giampaoli and Singer, 2009; Saville and Herring, 2009; Sinha, 2009; Wiencierz et al., 2011; Drikvandi et al., 2013; Qu et al., 2013; Wood, 2013; Baey et al., 2019; Chen et al., 2019). In addition to being of practical interest, this case is of theoretical interest because the parameter is a boundary point of the parameter set and, consequently, asymptotic distributions of common test statistics are non-standard. For example, the asymptotic distribution of the likelihood ratio test statistic for a variance equal to zero is a non-trivial mixture of chi-square distributions (Self and Liang, 1987; Geyer, 1994; Stram and Lee, 1994), whereas for a strictly positive variance it is a chi-square distribution with one degree of freedom. More generally, the asymptotic distributions of common test statistics under a sequence of parameters tending to a boundary point as the sample size increases, can be different depending on the rate of that convergence (Rotnitzky et al., 2000; Bottai, 2003). While this need not be an issue when testing a point null hypothesis, it complicates more ambitious inference: coverage probabilities of confidence regions obtained by inverting such test statistics often depend substantially on how close to the boundary the true parameter is, leading to unreliable inference. We address this using a connection between boundary points and points where the Fisher information matrix is singular, which we call critical points. More specifically, we show many boundary points of interest are also critical points and use this to construct confidence regions that (i) have asymptotically correct uniform coverage probability, (ii) have empirical coverage close to nominal in simulations, and (iii) are straightforward to implement for many mixed models, including the ubiquitous linear mixed model. We know of no other confidence regions with properties (i) and (ii) in the settings we consider.

To be more precise about the connections between boundary points, critical points, and mixed models, suppose a parameter $\theta \in \mathbb{R}$ scales a random effect with mean zero and unit variance in a mixed model, implying θ^2 is a variance. For example, θ can be the coefficient of a random effect in a linear predictor in a generalized linear mixed model. If the random effect has a distribution asymmetric around zero, inference on both the sign and magnitude of θ may be possible, in which case $\theta = 0$ is not a boundary point. In other settings the sign is unidentifiable and inference on $\theta \geq 0$ and θ^2 essentially equivalent; $\theta = 0$ is a boundary point. Either way, we will show that in quite general mixed models $\theta = 0$ is a critical point. Similarly, when θ is a vector of parameters whose jth element θ_j is a scale parameter, $\theta = \theta_*$ is often a critical point if $\theta_{*j} = 0$. Whether in a mixed model or not, inference near critical points is known to be difficult: the likelihood ratio test statistic and the maximum likelihood estimator behave quite differently than under classical conditions (Rotnitzky et al., 2000) and confidence regions obtained by inverting common test statistics such as the Wald, likelihood ratio, and score standardized by observed information have incorrect coverage probabilities (Bottai, 2003). By contrast, we show that, under regularity conditions, (i) the score test statistic standardized by expected Fisher information has a continuous extension at critical points and, when inverted, (ii) that test statistic gives a confidence region with asymptotically correct uniform coverage probability on compact sets. That is, the confidence region $\mathcal{R}_n(\alpha)$ based on n observations with nominal level $(1 - \alpha) \in (0, 1)$ satisfies, for any compact subset C of the parameter set,

$$\lim_{n\to\infty} \inf_{\theta\in C} \mathsf{P}_\theta\{\theta \in \mathcal{R}_n(\alpha)\} = 1 - \alpha, \tag{1}$$

where the subscript θ on P indicates the data on which $\mathcal{R}_n$ is based have the distribution indexed by θ. Importantly, C can include boundary and critical points. It is an immediate corollary that the test rejecting a null hypothesis $\theta = \theta_0$ when $\theta_0 \notin \mathcal{R}_n(\alpha)$ has asymptotic size α for any θ_0. These results apply to but are not restricted to mixed models. Moreover, in contrast to many methods for testing variance parameters in mixed models, our in general does not require the implementation of simulation algorithms or computing the maximum likelihood estimator, which can be complicated in non-linear mixed models.

The connection between singular information and boundary settings has been noticed previously (Cox and Hinkley, 2000; Chesher, 1984; Lee and Chesher, 1986), but results similar to ours have only been obtained for settings with a single scalar parameter (Bottai, 2003). We recover those results as special cases, and under weaker conditions. Asymptotic properties of maximum likelihood estimators and likelihood ratio test statistics have been established for the special case where the rank of the Fisher information matrix is one less than full (Rotnitzky et al., 2000), but confidence regions were not considered. Notably, our theory does not require the information matrix to have a particular rank and, indeed, we will see that in mixed models the rank is often full minus the number of scale parameters equal to zero.

We end this section with a simple example that illustrates how critical points often appear in mixed models. After the example, we give additional background and develop theory in Section 2. In Section 3 we discuss the application to mixed models and verify the conditions of the theory from Section 2 in two such models. Section 4 presents simulation results, Section 5 contains a data example, and Section 6 concludes.

Example 1. Suppose, for $i = 1, \dots, n$ and $j = 1, \dots, r$,

$$Y_{i,j} = \theta W_i + E_{i,j},$$

where $\theta \in [0, \infty)$ and all W_i and $E_{i,j}$ are independent standard normal random variables. For example, r can be the number of observations in a cluster, n the number of clusters, and the random effect W_i used to model heterogeneity between clusters or dependence between observations in the same cluster. The $Y_i = [Y_{i1}, \dots, Y_{ir}]^\mathsf{T}$, $i = 1, \dots, n$, are independent and multivariate normally distributed with mean zero and common covariance matrix $\Sigma(\theta) = \theta^2 1_r 1_r^\mathsf{T} + I_r$, where 1_r is an r-vector of ones and I_r the $r \times r$ identity matrix. With some algebra (Appendix B), one can show that the log-likelihood for one observation $y_i \in \mathbb{R}^r$ is

$$\log f_\theta(y_i) = -\frac{1}{2}\log(1 + \theta^2 r) - \frac{1}{2}\left\{ y_i^\mathsf{T} y_i - (y_i^\mathsf{T} 1_r)^2 \theta^2 / (1 + r\theta^2) \right\}.$$

Differentiating with respect to θ gives the score function for one observation:

$$s(\theta, y_i) = -\frac{r\theta}{1 + r\theta^2} + (y_i^\mathsf{T} 1_r)^2 \frac{\theta}{(1 + r\theta^2)^2}.$$

At $\theta = 0$, this score is zero for any $y_i \in \mathbb{R}^r$ and, hence, the Fisher information is zero; that is, $\theta = 0$ is a critical point. There are no other critical points because the second term of $s(\theta, Y_i)$, $Y_i \sim f_\theta$, has positive variance when $\theta \neq 0$.

Figure 1 shows two (pseudo) randomly generated realizations of the log-likelihood in this example. For one dataset the critical point is a global maximizer and for the other a local minimizer. One can show that if the true θ is small, both types of outcomes have probability approximately 1/2. In particular, the score always vanishes at $\theta = 0$ and the maximum likelihood estimator for θ is zero with probability approximately 1/2. The maximum likelihood estimator's mass at zero gives some intuition for why confidence regions that directly or indirectly use asymptotic normality of that estimator can have poor coverage properties near the critical point (see Rotnitzky et al., 2000, and Bottai, 2003, for details). In this example, the critical point is at the boundary since we assumed $\theta \geq 0$ for identifiability, but $\theta = 0$ would still be a critical point if the W_i had an asymmetric distribution and the sign of θ were identifiable.

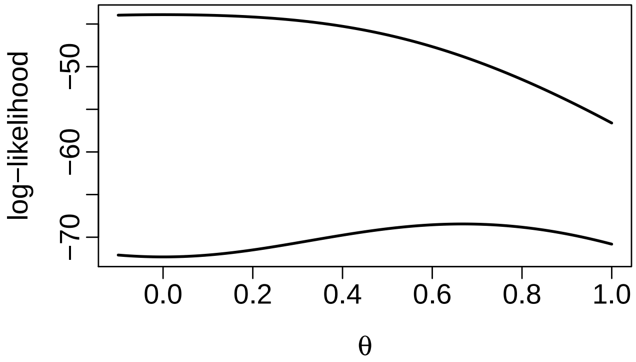

Figure 1: Log-likelihoods for two independent samples of $n = 100$ independent observations each, generated with $\theta = 1/\sqrt{10}$ and $r = 1$

2 Inference near critical points

2.1 Definitions and assumptions

Suppose a random vector $Y \in \mathbb{R}^r$ has density f_θ against a dominating measure γ, for θ in a parameter set $\Theta \subseteq \mathbb{R}^d$. In our motivating examples, θ is the parameter vector in a mixed model but there are other potential applications for the theory in this section. For example, Azzalini and Capitanio (2014, Section 3.1.3) consider a three-parameter multivariate skew-normal distribution with singular Fisher information. Let $s(\theta, y) = \nabla_\theta \log f_\theta(y) \in \mathbb{R}^d$ be the score function evaluated at (θ, y). The (expected) Fisher information (matrix) is $\mathcal{I}(\theta) =$

$\mathrm{cov}_\theta\{s(\theta, Y)\} \in \mathbb{R}^{d\times d}$. The subscript on cov_θ means Y has the distribution indexed by θ; we write $Y \sim f_\theta$. Focus will be on inference near points where the Fisher information is singular.

Definition 1. We say a point $\theta \in \Theta$ is critical if $\mathcal{I}(\theta)$ is singular and non-critical otherwise.

As mentioned in the Introduction, common test statistics typically do not satisfy (1) for subsets of the parameter set including critical points. To address this, we will consider the score test statistic standardized by expected Fisher information. For non-critical θ and realizations $y_1, \dots, y_n$, the test statistic is

$$T_n(\theta) = T_n(\theta; y_1, \dots, y_n) = \frac{1}{n}\left\{\sum_{i=1}^n s(\theta, y_i)^\mathsf{T}\right\}\mathcal{I}(\theta)^{-1}\left\{\sum_{i=1}^n s(\theta, y_i)\right\}. \tag{2}$$

The right-hand side of (2) is undefined at critical θ since $\mathcal{I}(\theta)^{-1}$ does not exist there. When θ is a scalar parameter, a continuous extension of T_n is, under regularity conditions (Bottai, 2003), equal to (2) at non-critical θ and at critical θ equal to

$$\frac{1}{n}\left\{\sum_{i=1}^n \frac{\partial^2 \log f_\theta(y_i)}{\partial\theta^2}\right\}^2 \mathrm{var}_\theta\left\{\frac{\partial^2 \log f_\theta(Y)}{\partial\theta^2}\right\}^{-1}.$$

That is, at critical points the test statistic is based on the second derivative of the log-likelihood. In what follows, we use the same notation for T_n and its continuous extension whenever the latter has been shown to exist. Now, our interest is twofold, namely conditions that ensure (i) T_n has a continuous extension to critical points when $d \geq 1$ and (ii) confidence regions obtained by inverting that extension satisfy (1). Both (i) and (ii) are substantially more complicated when $d \geq 1$ than when $d = 1$ since the eigenvectors and the rank of the Fisher information matrix become important.

We will use the following assumptions.

Assumption 1. The support $\mathcal{Y} = \{y : f_\theta(y) > 0\}$ is the same for every $\theta \in \Theta$.

Assumption 2. For every $\theta \in \Theta$, there is an open ball $B \subseteq \mathbb{R}^d$ around θ on which, for every $y \in \mathcal{Y}$, partial derivatives of $\log f_\theta(y)$ with respect to elements of θ, of an order $k \geq 2$ to be specified, (i) exist; (ii) are jointly continuous in (θ, y); and (iii) satisfy, for some $\delta > 0$,

$$\sup_{\theta\in B\cap\Theta, \tilde\theta\in B} \int |\partial^{(l)} \log f_{\tilde\theta}(y)/\partial\theta^{(l)}|^{2+\delta} f_\theta(y)\gamma(\mathrm{d}y) < \infty \quad (l = 1, \dots, k, j = 1, \dots, d),$$

where $\partial^{(l)} \log f_{\tilde{\theta}}(y)/\partial\theta^{(l)}$ denotes an arbitrary partial derivative of order l of $\theta \mapsto \log f_\theta(y)$ evaluated at $\tilde{\theta}$.

The open ball in Assumption 2 can include points not in Θ. In such cases, the assumption should be understood as saying there is an extension of s to $(\Theta \cup B) \times \mathcal{Y}$ satisfying the outlined conditions. Similarly, the continuity in (ii) is on $B \times \mathcal{Y}$; continuity need not hold on $B \times \mathbb{R}^r$ unless $\mathcal{Y} = \mathbb{R}^d$. We have sacrificed some generality for clarity in that, as will be clear later, the moment condition (iii) can be weakened to apply to only certain partial derivatives of the order k to be specified.

Assumption 3. For every $\theta \in \Theta$, orthonormal eigenvectors $\{v_{\theta 1}, \dots, v_{\theta d}\}$ of $\mathcal{I}(\theta)$ can be selected so that if $\mathcal{I}(\theta)v_{\theta j} = 0$, then $\mathcal{I}(\tilde{\theta})v_{\theta j} = 0$ for all $\tilde{\theta} \in \Theta$ with $\tilde{\theta}_j = \theta_j$.

Loosely speaking, Assumption 3 says an eigenvector of the Fisher information with vanishing eigenvalue corresponds to an element of the parameter vector being equal to a particular value; it does not matter what the other elements of the parameter vector are.

Definition 2. We say that an element θ_j of θ is critical at θ_{*j}, with corresponding critical (eigen-)vector v, if $\mathcal{I}(\theta)v = 0$ at every θ with $\theta_j = \theta_{*j}$.

In general, an eigenvector of the Fisher information with vanishing eigenvalue need not be a critical vector. For example, if v_1 and v_2 are critical vectors with corresponding critical elements $\theta_1 = \theta_{*1}$ and $\theta_2 = \theta_{*2}$, then any linear combination of v_1 and v_2 is also an eigenvector with vanishing eigenvalue at every θ where both $\theta_1 = \theta_{*1}$ and $\theta_2 = \theta_{*2}$; but there is not a corresponding critical element. An implication of Assumption 3 is that the rank of $\mathcal{I}(\theta)$ is d minus the number of critical elements. We will see that in mixed models it is often the case that scale parameters at zero are critical elements and standard basis vectors are the corresponding critical vectors.

Assumption 3 is sensitive to parameterization. For example, in the model considered by Azzalini and Capitanio (2014, Section 3.1), Assumption 3 does not hold in the first parameterization discussed by the authors but can be made to hold by a simple reparameterization. Because the test statistic (2) is invariant under differentiable reparameterizations with full rank Jacobian, it suffices to verify the conditions in one parameterization for the results to apply more generally.

Assumption 4. For every θ and $\{v_{\theta 1}, \dots, v_{\theta d}\}$ satisfying Assumption 3, there are $k = k(j) \in \{0, 1, \dots\}$ such that, with $Y \sim f_\theta$, the random vector

$$\tilde{s}(\theta, Y) = \left[v_{\theta 1}^{\mathsf{T}} \frac{\partial^{k(1)}}{\partial \theta_1^{k(1)}} s(\theta, Y), \dots, v_{\theta d}^{\mathsf{T}} \frac{\partial^{k(d)}}{\partial \theta_d^{k(d)}} s(\theta, Y) \right]$$

has mean zero and positive definite covariance matrix, where $k(j)$ can depend on θ_j but no other elements of θ.

The k in Assumption 4 should be the same as those in Assumption 2. Assumption 4 is a multidimensional version of the requirement that the second derivative of the log-likelihood has positive variance at critical points, as required by Bottai (2003). Indeed, Assumption 4 specializes to that requirement if $d = 1$, $k = 0$ at non-critical points, and $k = 1$ at critical points. The assumption can loosely be understood as saying that if a critical element θ_j with corresponding critical vector v_j is moved slightly, then v_j is no longer an eigenvector of the Fisher information with vanishing eigenvalue. At non-critical θ,

$$\tilde{s}(\theta, Y) = [v_{\theta 1}, \dots, v_{\theta d}]^{\mathsf{T}} s(\theta, Y),$$

whose covariance matrix, by definition of the $v_{\theta j}$, is a diagonal matrix with the eigenvalues of $\mathcal{I}(\theta)$ on the diagonal. Since these are positive by definition at non-critical points, Assumption 4 holds automatically at non-critical points.

Assumption 5. The non-critical elements are dense in Θ; that is, for any $\theta \in \Theta$, there exists a sequence $\{\theta_m\} \in \Theta$ of non-critical points tending to θ.

2.2 Continuous extension

Our purpose in this section is to prove the following theorem.

Theorem 2.1. *Under Assumptions 1 – 5, $T_n(\cdot; \cdot)$ has a continuous extension on $\Theta \times \mathcal{Y}^n$ for any $n \geq 1$.*

Observe that the continuity of T_n in Theorem 2.1 is jointly in the parameter and data. Before giving a proof, we state and discuss some intermediate results used in that proof. The first is a lemma which, loosely speaking, says that if the non-critical points are dense, then it is enough to establish continuity along sequences of non-critical points for it to hold more generally. The proof is in Appendix A.

Lemma 2.2. *Suppose Assumption 5 holds and that, for every $\theta \in \Theta$ and $(y_1, \ldots, y_n) \in \mathcal{Y}^n$, with n fixed, $\lim_{m\to\infty} T_n(\theta_m; y_{m1}, \ldots, y_{mn})$ exists and is the same for all sequences $\{\theta_m\} \in \Theta$ of non-critical points tending to θ and sequences $\{(y_{m1}, \ldots, y_{mn})\} \in \mathcal{Y}^n$ tending to $(y_1, \ldots, y_n)$; then the conclusion of Theorem 2.1 holds.*

In order to use Lemma 2.2 to prove Theorem 2.1, one must show that T_n converges along sequences of non-critical points tending to critical points. Evaluating the score at such sequences gives a sequence of score vectors whose covariance matrices are non-singular but tend to a singular limit. However, the following lemma says the score vectors can be scaled to tend to a limit with positive definite covariance matrix.

Lemma 2.3. *Suppose Assumptions 1 – 4 hold and let $\{\theta_m\} \in \Theta$ be a sequence of non-critical points tending to a $\theta \in \Theta$. Then there exist sequences of non-zero constants $\{a_{mj}\}$, $j = 1, \ldots, d$, such that, for any $\{y_m\} \in \mathcal{Y}$ tending to a $y \in \mathcal{Y}$ as $m \to \infty$,*

$$v_{\theta j}^{\top} s(\theta_m, y_m)/a_{mj} \to v_{\theta j}^{\top} \frac{\partial^k}{\partial \theta_j^k} s(\theta, y),$$

where $\{v_{\theta 1}, \ldots, v_{\theta d}\}$ and $k = k(j)$ are given by Assumption 4.

Proof. Since the limit point θ is fixed, denote $v_j = v_{\theta j}$ for simplicity. For an arbitrary j and all large enough m, Assumption 2 lets us apply Taylor's theorem with Lagrange-form remainder to the map $\theta_{mj} \mapsto v_j^{\top} s(\theta_m, y_m)$ to get

$$v_j^{\top} s(\theta_m, y_m) = \sum_{l=0}^{k-1} \frac{(\theta_{mj} - \theta_j)^l}{l!} \frac{v_j^{\top} \partial^l s(\theta_m^{(j)}, y_m)}{\partial \theta_j^l} + \frac{(\theta_{mj} - \theta_j)^k}{k!} \frac{v_j^{\top} \partial^k s(\tilde{\theta}_m^{(j)}, y_m)}{\partial \theta_j^k},$$

where $\theta_m^{(j)}$ and $\tilde{\theta}_m^{(j)}$ are θ_m with θ_{mj} replaced by, respectively, θ_j and a point between θ_{mj} and θ_j. By Assumption 3, $\theta_m^{(j)}$ is a critical point with critical vector v_j, so the first $k-1$ terms on right-hand side vanish for $y_m \in \mathcal{Y}$, giving

$$v_j^{\top} s(\theta_m, y_m) = \frac{(\theta_{mj} - \theta_j)^k}{k!} \frac{\partial^k v_j^{\top} s(\tilde{\theta}_m^{(j)}, y_m)}{\partial \theta_j^k}.$$

Setting $a_{mj} = (\theta_{mj} - \theta_j)^k / k!$, which is non-zero since θ_{mj} is non-critical, and using continuity of the partial derivatives given by Assumption 2 finishes the proof. □

The importance of controlling the behavior of the eigenvectors of the Fisher information matrix near critical points is highlighted by the proof of Lemma 2.3: the first $k-1$ terms in the Taylor expansion need not vanish if the eigenvectors depend on θ in a way violating Assumption 3. We are ready to prove Theorem 2.1.

Proof of Theorem 2.1. By Lemma 2.2 and definition of T_n, it suffices to show

$$\lim_{m\to\infty}\left\{s(\theta_m, y_{m1})^\mathsf{T}\mathcal{I}(\theta_m)^{-1}s(\theta_m, y_{m2})\right\}$$

exists and is the same for any non-critical $\{\theta_m\}$ tending to θ and $\{(y_{m1}, y_{m2})\}$ tending to (y_1, y_2). Let $A_m = \operatorname{diag}(a_{m1}, \dots, a_{md})$ be defined by the sequences of constants given by Lemma 2.3 and let $V = [v_{\theta 1}, \dots, v_{\theta d}] \in \mathbb{R}^{d\times d}$. Then $A_m V^\mathsf{T}$ is invertible and $s(\theta_m, y_{m1})^\mathsf{T}\mathcal{I}(\theta_m)^{-1}s(\theta_m, y_{m2})$ is equal to

$$\{A_m V^\mathsf{T} s(\theta_m, y_{m1})\}^\mathsf{T}\{A_m V^\mathsf{T}\mathcal{I}(\theta_m)A_m V\}^{-1}A_m V^\mathsf{T}s(\theta_m, y_{m2}).$$

Lemma 2.3 says, for the $\tilde{s}$ defined in Assumption 4, $A_m V^\mathsf{T} s(\theta_m, y_{m1}) \to \tilde{s}(\theta, y_1)$, and similarly for $y_{m2} \to y_2$. Next observe that $A_m V^\mathsf{T}\mathcal{I}(\theta_m)A_m V = \operatorname{cov}_{\theta_m}\{A_m V^\mathsf{T} s(\theta_m, Y)\}$; we will show this tends to $\operatorname{cov}_\theta\{\tilde{s}(\theta, Y)\}$. To that end, we first show $A_m V^\mathsf{T} s(\theta_m, Y_m) \rightsquigarrow \tilde{s}(\theta, Y)$, where $Y_m \sim f_{\theta_m}$ and $Y \sim f_\theta$. Since $\theta_m \to \theta$, $f_{\theta_m} \to f_\theta$ pointwise in y by continuity implied by Assumption 2, and hence $Y_m \to Y$ in total variation by Scheffe's theorem (Billingsley, 1995, Theorem 16.12), and hence also in distribution. To show the desired convergence, we may thus assume, by Skorokhod's representation theorem (Billingsley, 1999, Theorem 6.7), that $Y_m \to Y$ almost surely. But then $A_m V^\mathsf{T} s(\theta_m, Y_m) \to \tilde{s}(\theta, Y)$ almost surely by Lemma 2.3, which implies the desired convergence in distribution. It remains only to show the corresponding covariance matrices also converge. To this end, observe that, as in the proof of Lemma 2.3, the jth element of $A_m V^\mathsf{T} s(\theta_m, Y_m)$ is $v_{\theta j}^\mathsf{T}\partial^{k(j)} s(\tilde{\theta}_m^{(j)}, Y_m)/\partial\theta_j^{k(j)}$, and thus the (i, j)th element of the covariance matrix is

$$\mathbb{E}\left\{v_{\theta i}^\mathsf{T}\frac{\partial^{k(i)}}{\partial\theta_i^{k(i)}}s(\tilde{\theta}_m^{(i)}, Y_m)v_{\theta j}^\mathsf{T}\frac{\partial^{k(j)}}{\partial\theta_j^{k(j)}}s(\tilde{\theta}_m^{(j)}, Y_m)\right\}.$$

Cauchy–Schwartz's inequality (if $i \neq j$) and Assumption 2 imply the absolute value of the random variable in the expectation has bounded $(1+\delta/2)$th moment, and hence the sequence is uniformly integrable. It follows (Billingsley, 1995, Theorem 25.12) that the second moments of $A_m V^\mathsf{T} s(\theta_m, Y)$ converge under θ_m. Since its mean is zero by the first Bartlett identity, it follows, as desired, $\mathrm{cov}_{\theta_m}\{A_m V^\mathsf{T} s(\theta_m, Y)\} \to \mathrm{cov}_\theta\{\tilde{s}(\theta, Y)\}$. Now, since the last right-hand side is invertible by Assumption 4, the inverses of the left-hand side converge to the inverse of the right-hand side by continuity. Thus, we have proven that

$$s(\theta_m, y_{m1})^\mathsf{T} \mathcal{I}(\theta_m)^{-1} s(\theta_m, y_{m2}) \to \tilde{s}(\theta, y_1)^\mathsf{T}\, \mathrm{cov}_\theta\{\tilde{s}(\theta, Y)\}^{-1} \tilde{s}(\theta, y_2).$$

To complete the proof, observe that the right-hand side is indeed the same for every $\{\theta_m\}$ and $\{(y_{m1}, \ldots, y_{mn})\}$ since the left-hand side is the same for every choice of V. □

2.3 Asymptotic uniform coverage probability

We now turn to confidence regions obtained by inverting the continuous extension T_n. Specifically, for $\alpha \in (0, 1)$, define

$$\mathcal{R}_n(\alpha) = \{\theta \in \Theta : T_n(\theta) \leq q_{d,1-\alpha}\}, \tag{3}$$

where $q_{d,1-\alpha}$ is the $(1-\alpha)$th quantile of the chi-square distribution with d degrees of freedom. We have the following main result of the section.

Theorem 2.4. *Under Assumptions 1 – 5, the confidence region $\mathcal{R}_n(\alpha)$ in* (3) *has asymptotically correct uniform coverage probability on compact sets; that is, it satisfies* (1).

We need some intermediate results before proving Theorem 2.4. The following two lemmas let us focus on convergence along sequence of non-critical points as in the previous section, but now taking the stochastic properties of the data into account.

Lemma 2.5. *Equation* (1) *holds if, for every convergent sequence $\{\theta_n\} \in \Theta$ as $n \to \infty$,*

$$T_n(\theta_n; Y_{n1}, \ldots, Y_{nn}) \rightsquigarrow \chi^2_d, \tag{4}$$

where $Y_{n1}, \ldots, Y_{nn}$ are independent with common density f_{θ_n}.

The proof of Lemma 2.5 essentially amounts to showing continuous convergence implies (in fact, is equivalent to) uniform convergence on compact sets. It is provided in Appendix A for completeness.

Lemma 2.6. *If Assumptions 1 – 5 hold and* (4) *holds for every convergent sequence* $\{\theta_n\} \in \Theta$ *of non-critical points, then* (4) *holds for any convergent sequence in* Θ*.*

Proof. Let F_n denote the cumulative distribution function of $T_n(\theta_n; Y_{n1}, \dots, Y_{nn})$ and let F denote that of χ^2_d. Assumption 5 says that, for every fixed n, we can pick a sequence of non-critical points $\{\tilde{\theta}_{mn}\}$ tending to θ_n. Assumptions 1 – 5 ensure Theorem 2.1 holds and this implies, essentially by the continuous mapping theorem (see Lemma A.1),

$$T_n(\tilde{\theta}_{mn}; Y_{mn1}, \dots, Y_{mnn}) \rightsquigarrow T_n(\theta_n; Y_{n1}, \dots, Y_{nn}), \quad m \to \infty,$$

where $Y_{mni} \sim f_{\tilde{\theta}_{mn}}$. Thus, the corresponding cumulative distribution function $\tilde{F}_{mn}$ tends to F_n at every point of continuity of F_n. Let D_n be the set of discontinuities of F_n and $D = \cup_n D_n$. Since any cumulative distribution function has at most countably many discontinuities, D is countable as a countable union of countable sets. Now for any $t \in \mathbb{R} \setminus D$, we can pick, for every n, an $m = m(n)$ large enough that $\|\tilde{\theta}_{mn} - \theta_n\| \leq 1/n$ and $|\tilde{F}_{mn}(t) - F_n(t)| \leq 1/n$; for simplicity denote a $\tilde{F}_{mn}$ satisfying this by $\tilde{F}_n$. We now have by the triangle inequality,

$$|F_n(t) - F(t)| \leq 1/n + |\tilde{F}_n(t) - F(t)|,$$

which tends to zero by the assumption that (4) holds along sequences of non-critical points since F is continuous; in particular, t is a point of continuity of F. The proof is completed by observing that, since D is countable, $\mathbb{R} \setminus D$ is dense in $\mathbb{R}$ and hence the convergence in fact holds at every $t \in \mathbb{R}$ (Fristedt and Gray, 2013, Proposition 2, Chapter 14). □

We are ready to prove Theorem 2.4

Proof of Theorem 2.4. By Lemma 2.6, it suffices to consider an arbitrary sequence $\{\theta_n\}$ of non-critical points tending to a $\theta \in \Theta$. The proof idea is to first use Assumption 2 and arguments similar to those in the proof of Lemma 2.3 to show that Lyapunov's condition (Billingsley, 1995, Theorem 27.3) holds for (any linear combination of) the scaled score

$$U_n := n^{-1/2} \sum_{i=1}^{n} A_n V^{\mathsf{T}} s(\theta_n, Y_{ni}),$$

where $V = [v_1, \dots, v_d] = [v_{\theta 1}, \dots, v_{\theta d}] \in \mathbb{R}^{d \times d}$ a matrix whose columns satisfy Assumption 3 at θ and $A_n = \mathrm{diag}(a_{n1}, \dots, a_{nd})$ is a scaling matrix defined by the $\{a_{nj}\}$ given by Lemma 2.3

(with the index $m = n$), and $Y_{n1}, \dots, Y_{nn}$ are independent with common density f_{θ_n}. Then, the result follows by Slutsky's theorem since the covariance matrix $\mathrm{cov}_\theta\{A_n V^\top s(\theta, Y)\}$ has a positive definite limit by the arguments in the proof of Theorem 2.1. We provide the details in Appendix A. □

Theorems 2.1 and 2.4 have the following corollary which recovers a result of Bottai (2003) but with several conditions weakened. The proof is a straightforward verification of Assumptions 2 – 5 and hence omitted.

Corollary 2.7. *If $d = 1$, Assumption 1 holds, and for every $\theta \in \Theta$ there is an open ball $B \subseteq \mathbb{R}$ including θ on which $h(\theta, y) = \partial^2 \log f_\theta(y)/\partial\theta^2$ satisfies (i) $\int |h(\tilde\theta, y)|^{2+\delta} f_\theta(y)\, \gamma(\mathrm{d}y) < \infty$ uniformly in $(\theta, \tilde\theta) \in (B \cap \Theta) \times B$ for some $\delta > 0$ and (ii) $\mathrm{var}_\theta\{h(\theta, Y)\} > 0$ for every critical θ; then the conclusions of Theorems 2.1 and 2.4 hold.*

We end this section by illustrating the introduced ideas in the example from the introduction. More complicated mixed models are considered in the next section.

Example 1 (continued)**.** Recall that the $Y_i \in \mathbb{R}^r$, $i = 1, \dots, n$, are independent and multivariate normally distributed with mean 0 and covariance matrix $\Sigma(\theta) = \theta^2 1_r 1_t^\top + I_r$, and that the score function for one observation is

$$s(\theta, y_i) = -\frac{r\theta}{1 + r\theta^2} + (y_i^\top 1_r)^2 \frac{\theta}{(1 + r\theta^2)^2} = \frac{\theta}{1 + r\theta^2}\left(-r + \frac{(y_i^\top 1_r)^2}{1 + r\theta^2}\right)$$

Since Y_i has positive variance for all θ, the score function is almost surely equal to zero if and only if $\theta = 0$, verifying Assumptions 3 and 5. Assumption 1 holds with γ being Lebesgue measure on $\mathbb{R}^r$. To verify Assumption 4, observe that at $\theta = 0$ the second derivative of the log-likelihood is

$$\lim_{\theta \to 0} \frac{s(\theta, y_i) - s(0, y_i)}{\theta} = \lim_{\theta \to 0} \left[\frac{1}{1 + r\theta^2}\left(-r + \frac{(y_i^\top 1_r)^2}{1 + r\theta^2}\right)\right] = -r + (y_i^\top 1_r)^2,$$

which has positive variance. Thus, Assumption 4 holds at $\theta = 0$ with $k = 1$ and at $\theta \neq 0$ with $k = 0$. It is straightforward to verify Assumption 2 and hence conclude Theorems 2.1 and 2.4 apply. For additional insight, we also provide a more direct argument for why the score test standardized by expected information works in this example while other common test statistics do not. At $\theta \neq 0$ (see Appendix B),

$$T_n(\theta; Y_1, \dots, Y_n) = \frac{1}{2rn}\left\{-rn + \sum_{i=1}^{n}(Y_i^{\mathsf{T}} 1_r)^2/(1+r\theta^2)\right\}^2 \sim \frac{(-rn + r\chi_n^2)^2}{2rn}.$$

It is immediate from the middle expression that $T_n(\cdot\,;\cdot)$ has a continuous extension on $[0,\infty) \times \mathbb{R}^{nr}$. Essentially, standardizing by expected information cancels the leading factor $\theta/(1+r\theta^2)$ in the expression for $s(\theta, y_i)$, which was the reason for the singularity at $\theta = 0$. Notably, this cancellation does not happen if one instead standardizes by observed information. The last expression shows that, in this example, the distribution of the proposed test statistic is in fact independent of the parameter, and hence it is almost immediate that (4) and, hence, Theorem 2.4 hold; writing χ_n^2 as the sum of n independent χ_1^2 and an appeal to the classical central limit theorem is all that is needed. To emphasize the fact that other common test statistics do not enjoy the same asymptotic properties, we show in Theorem A.2 in Appendix A that the asymptotic distribution of the score test statistic standardized by observed information evaluated at the true θ_n, is different depending on how $\{\theta_n\}$ tends to 0.

3 Inference near critical points in mixed models

3.1 Scale parameters at zero

Suppose $Y \in \mathbb{R}^r$ has conditional density $f_\theta(y \mid w)$ against γ given a vector of random effects $W \in \mathbb{R}^q$ whose elements are independent with mean zero, unit variance, and joint distribution ν. We partition the parameter vector as $\theta = (\lambda, \psi) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ and assume, for some parameter matrix $\Lambda = \Lambda(\lambda) \in \mathbb{R}^{q \times q}$ and known $h : \mathcal{Y} \times \mathbb{R}^{d_2} \times \mathbb{R}^q \to [0, \infty)$,

$$f_\theta(y \mid w) = h(y, \psi, \Lambda(\lambda)w).$$

We call λ a scale parameter since it determines the matrix $\Lambda(\lambda)$ scaling W. The vector ψ includes any other parameters in the distribution of $Y \mid W$. With these assumptions, the marginal distribution of Y has density against γ given by

$$f_\theta(y) = \int f_\theta(y \mid w)\, \nu(\mathrm{d}w) = \int h(y, \psi, \Lambda w)\, \nu(\mathrm{d}w). \tag{5}$$

For example, a generalized linear mixed model with linear predictor $X\psi + ZU$, where $U = \Lambda W \sim \mathcal{N}(0, \Lambda\Lambda^{\mathsf{T}})$, and $X \in \mathbb{R}^{r \times d_2}$ and $Z \in \mathbb{R}^{r \times q}$ are design matrices, satisfies (5) for an appropriate choice of h.

In practice it is often possible to define ψ, λ, Λ and h so that any critical points are so because the leading $d_1 \times d_1$ block of $\mathcal{I}(\theta)$ is singular. That is, all critical points are due to linear combinations of the scores for the scale parameters being equal to zero almost surely. To investigate when the latter can happen, let $\nabla_3 h(y, \psi, \Lambda w)$ denote the gradient of $h(y, \psi, \cdot)$ evaluated at Λw, the subscript 3 indicating derivatives with respect to the third argument. Then, assuming the derivatives exist and can be moved inside the integral, the jth element of the score function, $j = 1, \dots, d_1$, is

$$s_j(\theta, y) = \frac{1}{f_\theta(y)} \int \frac{\partial}{\partial \lambda_j} f_\theta(y)\, \gamma(\mathrm{d}y) = \frac{1}{f_\theta(y)} \int \{\nabla_3 h(y, \psi, \Lambda w)\}^{\mathsf{T}} \frac{\partial \Lambda(\lambda)}{\partial \lambda_j} w\, \nu(\mathrm{d}w). \tag{6}$$

The following result gives a sufficient condition for when linear combinations of these scores are equal to zero almost surely when evaluated at $y = Y \sim f_\theta$.

Proposition 3.1. *If* (6) *holds at* $\theta \in \Theta$ *and for a* $v = [v_1, \dots, v_{d_1}, 0, \dots, 0]^{\mathsf{T}} \in \mathbb{R}^d$ *it holds that*

$$\sum_{j=1}^{d_1} v_j \frac{\partial \Lambda(\lambda)}{\partial \lambda_j} W \quad \textit{and} \quad \Lambda(\lambda) W$$

are independent under θ*; then* $\mathcal{I}(\theta) v = 0$.

Proof. It suffices to show that $v^{\mathsf{T}} s(\theta, y) = 0$ for all $y \in \mathcal{Y}$. We have

$$v^{\mathsf{T}} s(\theta, y) = \frac{1}{f_\theta(y)} \int \{\nabla_3 h(y, \psi, \Lambda w)\}^{\mathsf{T}} \sum_{j=1}^{d_1} v_j \frac{\partial \Lambda(\lambda)}{\partial \lambda_j} w\, \nu(\mathrm{d}w).$$

By the assumptions, the integral is the expectation of the inner product of two independent random vectors. Thus, since (measurable) functions, and coordinate projections in particular, of independent random variables are independent, we get that the integral is

$$\int \{\nabla_3 h(y, \psi, \Lambda w)\}^{\mathsf{T}}\, \nu(\mathrm{d}w) \int \sum_{j=1}^{d_1} v_j \frac{\partial \Lambda(\lambda)}{\partial \lambda_j} w\, \nu(\mathrm{d}w),$$

which is equal to zero since $\int w\, \nu(\mathrm{d}w) = 0$, and this completes the proof. □

If $v \in \mathbb{R}^d \setminus \{0\}$ and $\theta \in \Theta$ satisfy Proposition 3.1, then v is an eigenvector of the Fisher information matrix with vanishing eigenvalue, and hence θ is a critical point. The condition that $\sum_{j=1}^{d_1} v_j\{\partial\Lambda/\partial\lambda_j\}W$ and ΛW are independent random variables is not a necessary condition. We will see, however, that in practice it is often the case that all critical points are identified by Proposition 3.1.

In what what follows, to facilitate further analysis, we assume Λ is diagonal. Specifically, we assume every diagonal element of Λ is one of the λ_j. Then the λ_j are scale parameters in the usual sense and

$$\Lambda(\lambda) = \operatorname{diag}(\lambda_{j(1)}, \dots, \lambda_{j(q)}), \tag{7}$$

where $\lambda_{j(k)}$ means the λ_j in the kth diagonal element of $\Lambda(\lambda)$. This assumption is common in practice and is less restrictive than it may first seem: it allows for the possibility that w affects the conditional density $f_\theta(y \mid w)$ through $H\Lambda w$ for some $H \in \mathbb{R}^{q\times q}$, which may depend on ψ. Then, $U = H\Lambda W$ can be viewed as a vector of dependent random effects whose scales are determined by λ and whose dependence structure is determined by H. In particular, by taking H to be an orthogonal matrix the λ_j^2 are the eigenvalues of $\operatorname{cov}_\theta(U) = H\Lambda(\lambda)^2 H^\mathsf{T}$.

With (7), Proposition 3.1 has the following corollary which says standard basis vectors are critical vectors for scale parameters at zero.

Corollary 3.2. *If* (6) *holds at $\theta \in \Theta$, $\Lambda(\lambda)$ satisfies* (7)*, and $\theta_j = \lambda_j = 0$; then $\mathcal{I}(\theta)e_j = 0$, where e_j is the jth standard basis vector in $\mathbb{R}^d$.*

Proof. With (7), $\partial\Lambda(\lambda)/\partial\lambda_j$ is a diagonal matrix whose kth diagonal element is 1 if the corresponding element of $\Lambda(\lambda)$ is λ_j, and zero otherwise. Thus, at θ such that $\lambda_j = 0$, $\{\partial\lambda(\lambda)/\partial\lambda_j\}W$ is a function of the W_k scaled by λ_j, and $\Lambda(\lambda)W$ a function of the W_k not scaled by λ_j. Thus, Proposition 3.1 is satisfied with $v = e_j$. □

We illustrate the wide applicability of Corollary 3.2 to popular mixed models using another example.

Example 2 (Generalized linear mixed model)**.** Let $X \in \mathbb{R}^{r\times d_2}$ and $Z \in \mathbb{R}^{r\times q}$ be design matrices and suppose

$$f_\theta(y \mid w) = h(y, \psi, \Lambda w) = \exp\left\{y^\mathsf{T}(X\psi + Z\Lambda w) - c(X\psi + Z\Lambda w)\right\},$$

where $c : \mathbb{R}^r \to \mathbb{R}$ is an elementwise cumulant function (see e.g. McCulloch et al., 2008, for definitions); for example, the jth element of $\nabla c(X\psi + Z\Lambda w)$, the gradient of c evaluated at the linear predictor, is the conditional mean of the jth response given w. Assume (7) and, for $j = 1, \dots, d_1$, let $[j]$ denote the set of $k \in \{1, \dots, q\}$ such that $\Lambda_{kk} = \lambda_j$. Then,

$$\partial Z\Lambda(\lambda)w/\partial\lambda_j = \sum_{k\in[j]} Z^k w_k,$$

where Z^k is the kth column of Z. Hence, assuming differentiation under the integral is permissible,

$$s_j(\theta, y) = \frac{1}{f_\theta(y)} \int f_\theta(y \mid w)\{y - \nabla c(X\psi + Z\Lambda w)\}^\mathsf{T} \sum_{k\in[j]} Z^k w_k \, \nu(\mathrm{d}w).$$

Corollary 3.2 suggests $s_j(\theta, y)$ is zero for any $y \in \mathcal{Y}$ and θ such that $\lambda_j = 0$, which can here be verified directly: the sum in the integral is a function of the w_k scaled by λ_j, while the remaining part of the integrand is a function the w_k not scaled by λ_j. Hence, the integral is

$$\int f_\psi(y \mid \Lambda w)[y - \nabla c(X\psi + Z\Lambda w)]^\mathsf{T} \, \nu(\mathrm{d}w) \int \sum_{k\in[j]} Z^k w_k \, \nu(\mathrm{d}w),$$

which is equal to zero for any $y \in \mathcal{Y}$ since the second integral is the expectation of a linear combination of the w_k, who have mean zero.

It is clear from Example 2 that critical points occur in many mixed models. Our theory in Section 2 says reliable inference is nevertheless possible under certain conditions. The following result will be helpful when verifying Assumptions 3 – 5 in examples.

Proposition 3.3. *If (i) the null space of $\mathcal{I}(\theta)$ is spanned by $\{e_j, j : \lambda_j = 0\}$, (ii) there exists a sequence $\{\theta_m\} = \{(\lambda_m, \psi)\} \in \Theta$ such that $\lambda_{mj} \neq 0$ and $\lambda_{mj} \to 0$ for all $j \in \{1, \dots, d_1\}$, and (iii) the random vector*

$$\bar{s}(\theta, Y) = \left[\frac{\partial^{k(1)+1} \log f_\theta(Y)}{\partial\theta^{k(1)+1}}, \dots, \frac{\partial^{k(d)+1} \log f_\theta(Y)}{\partial\theta^{k(d)+1}}\right]^\mathsf{T}, \quad Y \sim f_\theta,$$

where $k(j) = 1$ if $\theta_j = \lambda_j = 0$ and $k(j) = 0$ otherwise, has positive definite covariance matrix with finite entries; then Assumptions 3 – 5 hold.

The proof of Proposition 3.3 is in Appendix A. To summarize briefly, observe that a critical element $\theta_j = \lambda_j = 0$ with critical vector e_j is consistent with Assumption 3. Thus, a sufficient condition for Assumption 3 is that there are no critical points other than those induced by scale parameters at zero; that is, the null space of $\mathcal{I}(\theta)$ is spanned by $\{e_j, j : \lambda_j = 0\}$. Knowing that only standard basis vectors can be eigenvectors with vanishing eigenvalues simplifies things. When it holds, (ii) ensures Assumption 5 holds. Moreover, the elements of the vector $\tilde{s}(\theta, y)$ in Assumption 4 then specialize to include no cross-partial derivatives of the log-likelihood, in which case positive definiteness of $\tilde{s}(\theta, Y)$ is equivalent to that of $\bar{s}(\theta, Y)$, which is the usual score function modified to include a second order partial derivative when the corresponding first order partial derivative is equal to zero almost surely.

In the following two sections, we verify the conditions of Theorems 2.1 and 2.4 in two mixed models. The first is an exponential mixed model for correlated, positive responses. It has one scale parameter, one fixed effect parameter, and a non-normal random effect. This example illustrates the calculations needed to verify our conditions in non-linear mixed effects models with non-standard random effect distributions. The second model is a quite general version of the linear mixed model with normally distributed random effects. It has general design matrices X and Z and hence several fixed and random effect parameters.

3.2 Exponential mixed model with uniform random effect

Suppose $Y \in \mathbb{R}^2$ has conditionally independent elements given $W \in \mathbb{R}$ with

$$f_\theta(y_j \mid w) = (\psi + \lambda w)\exp\{-y_j(\psi + \lambda w)\}\mathbb{I}(y_j \geq 0); \quad \theta = (\lambda, \psi) \in \Theta \subseteq \mathbb{R}^2. \tag{8}$$

That is, the distribution of $Y_j \mid W$ is exponential with mean $1/(\psi + \lambda W)$. We have selected $r = 2$ responses to simplify calculations, but $r \geq 2$ presents no fundamental difficulties. The specification clearly requires $\psi + \lambda W$ be positive almost surely. There are many potentially useful specifications satisfying this, but to be concrete suppose W is uniform on $(-\sqrt{3}, \sqrt{3})$ and the parameter set $\Theta = \{(\lambda, \psi) \in [0, \infty) \times \mathbb{R} : \psi > \sqrt{3}\lambda\}$. The log-likelihood for one observation is, ignoring additive constants,

$$\log f_\theta(y) = \log \int_{-\sqrt{3}}^{\sqrt{3}} f_\theta(y \mid w) \mathrm{d}w = \log \int_{-\sqrt{3}}^{\sqrt{3}} (\psi + \lambda w)^2 \exp\{-(\psi + \lambda w)y_\bullet\} \mathrm{d}w,$$

where $y_\bullet = y_1 + y_2$. The score for λ is

$$s_\lambda(\theta, y) = -\frac{1}{f_\theta(y)} \int f_\theta(y \mid w)\{y_\bullet - 2/(\psi + \lambda w)\} w \,\mathrm{d}w.$$

Lemma 3.4. *In the exponential mixed model* (8)*, for any* $\theta = (0, \psi) \in \Theta$*,* $\mathrm{var}_\theta\{s_\lambda(\theta, Y)\} = 0$ *and* $\mathrm{var}_\theta\{\partial^2 \log f_\theta(Y)/\partial\lambda^2\} > 0$*.*

Proof. Setting $\theta = (0, \psi)$ in the expression for $s_\lambda(\theta, y)$ and moving terms that do not depend on w outside the integral gives $s_\lambda(0, \psi, y) = -f_\theta(y)^{-1}\psi \exp(-\psi y_\bullet)\{y_\bullet - 2/(\psi)\} \int w\mathrm{d}w = 0$. Moreover, when $s_\lambda(\theta, y) = 0$,

$$\frac{\partial^2 \log f_\theta(y)}{\partial \lambda^2} = -\frac{1}{f_\theta(Y)} \int f_\theta(Y \mid w)[\{Y_\bullet - 2/(\psi + \lambda w)\}^2 - 2/(\psi + \lambda w)^2] w^2 \mathrm{d}w.$$

When $\lambda = 0$ this simplifies to $(Y_\bullet - 2/\psi)^2 - 2/\psi^2$, which has positive variance under $\theta = (0, \psi)$. □

Lemma 3.4 shows $\theta = (0, \psi)$ is a critical point for any ψ, agreeing with Corollary 3.2. It also shows the corresponding second derivative of the log-likelihood has positive variance if $\lambda = 0$, suggesting it may be possible to verify Assumption 4 with $k(1) = 1$ at θ with $\lambda = 0$ and $k(1) = 0$ elsewhere. The following lemma will be helpful to that end.

Lemma 3.5. *In the exponential mixed model* (8)*, the information matrix* $\mathcal{I}(\theta)$ *has rank one if* $\lambda = 0$ *and rank two otherwise.*

Proof. The score for ψ is $s_\psi(\theta, y) = -f_\theta(y)^{-1} \int f_\theta(y \mid w)\{y_\bullet - 2/(\psi + \lambda w)\} \,\mathrm{d}w$. Thus, when $\lambda = 0$, $s_\psi(\theta, Y) = Y_\bullet - 2/\psi$ which has positive variance. In conjunction with Lemma 3.4, this shows the rank of $\mathcal{I}(\theta)$ is one when $\lambda = 0$. Suppose $\lambda > 0$ and make the change of variables $s = (\psi + \lambda w)y_\bullet$ in the integral in the definition of the log-likelihood. Letting $G(s) = e^{-s}(s^2 + 2s + 2)$, which is an antiderivative of $g(s) = -s^2 e^{-s}$, gives

$$\log f_\theta(y) = -\log(\lambda) + \log\{G(\psi - \sqrt{3} y_\bullet \lambda) - G(\psi + \sqrt{3} y_\bullet \lambda)\}.$$

Differentiating with respect to λ and ψ, taking an arbitrary linear combination given by $(v_1, v_2) \in \mathbb{R}^2$, and observing that $Y_\bullet$ has support on $(0, \infty)$ shows $\mathcal{I}(\theta)v = 0$ only if

$$s \mapsto \frac{v_1 \lambda^{-1} s\{g(\psi - s) + g(\psi + s)\} - v_2\{g(\psi - s) - g(\psi + s)\}}{G(\psi - s) - G(\psi + s)}$$

is constant on $(0, \infty)$, possibly except on a Lebesgue null set. Verifying this map is indeed non-constant on a set of positive Lebesgue measure is routine so we omit the details. □

We are ready to verify Assumptions 1 – 5, giving the main result of the section.

Theorem 3.6. *The conclusions of Theorems 2.1 and 2.4 hold in the exponential mixed model* (8).

Proof. Assumption 1 holds with $\mathcal{Y} = (0, \infty)$. To verify Assumption 2, it suffices by Lemma 3.5 to find locally uniform bounds of $\int |s_\lambda(\tilde{\theta}, Y)|^3 f_\theta(y)\,\mathrm{d}y$ and $\int |s_\psi(\tilde{\theta}, Y)|^3 f_\theta(y)\,\mathrm{d}y$ around an abitrary $\theta \in \Theta$, and of $\int |\partial^2 \log f_{\tilde{\theta}}(y)/\partial\lambda^2|^3 f_\theta(y)\,\mathrm{d}y$ if $\lambda = 0$. For the former, observe that for any $k \geq 1$, by Jensen's inequality and using $f_\theta(y \mid w)/f_\theta(y) \propto f_\theta(w \mid y)$ since $f(w) \propto 1$,

$$\begin{aligned}\int |s_\lambda(\theta, y)|^k f_\theta(y)\mathrm{d}y &= \int \left| \int \{y_\bullet - 2/(\psi + \lambda w)\} w f_\theta(w \mid y)\,\mathrm{d}w \right|^k f_\theta(y)\,\mathrm{d}y \\ &\leq \int \int |\{y_\bullet - 2/(\psi + \lambda w)\} w|^k f_\theta(y, w)\,\mathrm{d}w\,\mathrm{d}y.\end{aligned}$$

Now for any $\theta \in \Theta$, since $\psi > \lambda\sqrt{3} \geq \lambda w$, we can find a ball around θ small enough that $1/(\psi + \lambda w) \leq M < \infty$, and therefore the last line in the last display is uniformly bounded on that ball if and only if $\mathbb{E}_\theta(Y_j^k)$ is. But on that ball, $\mathbb{E}_\theta(Y_j^k) = \mathbb{E}_\theta\{\mathbb{E}_\theta(Y_j^k \mid W)\} = k!\mathbb{E}\{1/(\psi + \lambda W)^k\} \leq k! M^k$. The other moment bounds can be handled similarly to show Assumption 2 holds; we omit the details. We use Proposition 3.3 to verify Assumptions 3 – 5. Lemma 3.5 shows condition (i) of that proposition and condition (ii) is trivial. To verify (iii), it suffices by Lemma 3.5 to show that, when $\lambda = 0$, $v_1 \partial^2 \log f_\theta(Y)/\partial\lambda^2 + v_2 s_\psi(\theta, Y)$ is constant almost surely f_θ only if $v_1 = v_2 = 0$. But up to additive non-stochastic terms, that linear combination is $v_1(Y_\bullet - 2/\psi)^2 + v_2 Y_\bullet$, which clearly has positive variance unless $v_1 = v_2 = 0$, so we are done. □

3.3 Linear mixed models

Let $Z \in \mathbb{R}^{r \times q}$ be a non-stochastic design matrix $X \in \mathbb{R}^{r \times d_2}$ a matrix of predictors whose distribution does not depend on θ, including non-stochastic predictors as a special case. Consider a linear mixed model which assumes, for some $\beta \in \mathbb{R}^{d_2}$ and Λ satisfying (7),

$$Y \mid X, W \sim \mathcal{N}(X\beta + Z\Lambda W, \sigma^2 I_r), \quad W \mid X \sim \mathcal{N}(0, I_r). \tag{9}$$

Assume also for simplicity that $\sigma^2 > 0$ is known. When $\sigma^2 = 1$, this model is a special case of the generalized linear mixed model in Example 2. To fit in the general framework from Section 2, we assume independent copies of (Y, X): $(Y_i, X_i) \in \mathbb{R}^r \times \mathbb{R}^{r \times d_2}$, $i = 1, \ldots, n$.

We will first obtain a reliable confidence region for $\theta = (\lambda, \beta) \in \Theta = [0, \infty)^{d_1} \times \mathbb{R}^{d_2}$ by verifying the conditions of Theorem 2.4. Then, we show how that result can be modified to give a reliable confidence region for λ only, with β estimated. The model implies the distribution of $Y \mid X$ is multivariate normal with mean $X\beta$ and covariance matrix

$$\Sigma = \Sigma(\lambda) = \sigma^2 I_r + \sum_{j=1}^{d_1} \lambda_j^2 H_j,$$

where $H_j = \sum_{k \in [j]} Z^k (Z^k)^\mathsf{T} \in \mathbb{R}^{r \times r}$ is the sum of outer products of columns of Z corresponding to random effects scaled by λ_j, $j = 1, \dots, d_1$. Because the distribution of X does not depend on θ, the full log-likelihood is, up to terms not depending on θ, the same as that for $Y \mid X$:

$$-\frac{1}{2} \log |\Sigma(\lambda)| - \frac{1}{2}(y_i - X_i\beta)^\mathsf{T} \Sigma(\lambda)^{-1}(y_i - X_i\beta).$$

Differentiating with respect to λ_j gives, for $j = 1, \dots, d_1$,

$$s_j(\theta, y_i, X_i) = \lambda_j \operatorname{tr} \left\{ \Sigma^{-1} H_j - \Sigma^{-1}(y_i - X_i\beta)(y - X_i\beta)^\mathsf{T} \Sigma^{-1} H_j \right\},$$

which is equal to zero when $\lambda_j = 0$. Thus, scale parameters at zero are critical points, agreeing with Corollary 3.2. We also have the following partial converse, essentially saying that only scale parameters can be critical elements. The proof is routine but provided in Appendix A for completeness. We use s_λ to denote the score for λ, that is, the first d_1 elements of s, and similarly with s_β; and $\underline{\mathrm{e}}(\cdot)$ and $\bar{\mathrm{e}}(\cdot)$ denote, respectively, the smallest largest eigenvalues of their matrix arguments.

Lemma 3.7. *The Fisher information $\mathcal{I}(\theta)$ in the linear mixed model* (9) *is block-diagonal and, if $\underline{\mathrm{e}}\{\mathbb{E}(X^\mathsf{T} X)\} > 0$, is singular if and only if its leading block $\mathcal{I}_\lambda(\theta) = \operatorname{cov}_\theta\{s_\lambda(\theta, Y, X)\}$ is.*

The following lemma gives a lower bound on the variance of any linear combination of the score for the scale parameters. Together with the previous lemma, it can be used to identify all critical points.

Lemma 3.8. *For any $v \in \mathbb{R}^{d_1}$, in the linear mixed model* (9),

$$\operatorname{var}_\theta\{v^\mathsf{T} s_\lambda(\theta, Y, X)\} \geq 2\, \bar{\mathrm{e}}(\Sigma)^{-2}\, \underline{\mathrm{e}}(Z^\mathsf{T} Z)^3 \max_{j \in \{1, \dots, q\}} (\lambda_j v_j)^2,$$

with equality if $\max_{j \in \{1, \dots, q\}} (\lambda_j v_j)^2 = 0$.

The last term on the right-hand side in Lemma 3.8 ensures that, as long as Z has full column rank, $\{e_j, j : \lambda_j = 0\}$ spans the null space of $\mathcal{I}(\theta)$. As observed following Proposition 3.3, this makes checking the assumptions in Section 2 easier. We are ready for the first main result of the section. The proof is in Appendix A.

Theorem 3.9. *If $\underline{\mathrm{e}}(Z^\mathsf{T} Z) > 0$, $\underline{\mathrm{e}}\{\mathbb{E}(X^\mathsf{T} X)\} > 0$, and $\mathbb{E}(\|X\|^{4+\delta}) < \infty$ for some $\delta > 0$, then Assumptions 1 – 5, and hence the conclusions of Theorems 2.1 and 2.4, hold in the linear mixed model* (9)*.*

Lastly in this section, we consider confidence regions for λ with β estimated, specifically regions obtained by inverting

$$T_n^\lambda(\lambda; \beta) = \frac{1}{n}\left\{\sum_{i=1}^n s_\lambda(\theta, Y_i, X_i)^\mathsf{T}\right\}\mathcal{I}_\lambda(\theta)^{-1}\left\{\sum_{i=1}^n s_\lambda(\theta, Y_i, X_i)\right\}.$$

For such regions to be practically useful, or feasible, β has to be known or estimated. Our next result says β can estimated by any square root n-consistent estimator without affecting the asymptotic coverage probability. To be more specific, for a given β, let

$$\mathcal{R}_n^\lambda(\alpha) = \{\lambda : T_n^\lambda(\lambda; \beta) \leq q_{d_1, 1-\alpha}\}$$

and let $\hat{\mathcal{R}}_n^\lambda(\alpha)$ be that confidence interval with an estimator $\hat\beta$ in place of β in T_n. We have the following result whose proof is in Appendix A.

Theorem 3.10. *Under the conditions of Theorem 3.9, the confidence region $\mathcal{R}_n^\lambda(\alpha)$ satisfies, for any compact $C \subseteq \Theta$ and $\alpha \in (0, 1)$,*

$$\lim_{n\to\infty} \inf_{\theta \in C} \mathsf{P}_\theta\left\{\lambda \in \mathcal{R}_n^\lambda(\alpha)\right\} = 1 - \alpha;$$

and if in addition $\sqrt{n}\|\hat\beta - \beta\| = O_\mathsf{P}(1)$ under any convergent $\{\theta_n\} \in \Theta$, then the same holds for $\hat{\mathcal{R}}_n^\lambda(\alpha)$.

In practice, the estimator $\hat\beta$ can be, for example, the least squares estimator

$$\hat\beta = \left(\sum_{i=1}^n X_i^\mathsf{T} X_i\right)^{-1} \sum_{i=1}^n X_i^\mathsf{T} Y_i.$$

This estimator is square root-n consistent under convergent $\{\theta_n\}$ since it is multivariate normally distributed given $(X_1, \ldots, X_n)$ with mean β_n and conditional covariance matrix

$$\left(\sum_{i=1}^{n} X_i^{\mathsf{T}} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} \Sigma_n X_i\right) \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} X_i\right)^{-1},$$

which can be shown to tend to a matrix of zeros using the law of large numbers.

4 Numerical experiments

We examine finite sample properties of the proposed method in a linear mixed model. Suppose that for $i = 1, \ldots n$ and $j = 1, \ldots, r$,

$$Y_{ij} = \beta_1 + \beta_2 X_{it} + U_{1i} + U_{2i} X_{ij} + E_{ij},$$

where $[U_{1i}, U_{2i}]^{\mathsf{T}} \sim \mathcal{N}\{0, \mathrm{diag}(\lambda_1^2, \lambda_2^2)\}$, independently for $i = 1, \ldots, n$ and independent of $E = [E_{1,1}, \ldots, E_{n,r}]^{\mathsf{T}} \sim \mathcal{N}(0, \sigma^2 I_{nr})$.

Our first simulations examine coverage probabilities of confidence regions for $(\sigma, \lambda_1, \lambda_2)$. Data were simulated with $\sigma = 1$ and, to examine behavior near critical points,

$$\lambda_1 = \lambda_2 \in \{0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}.$$

The true coefficient was set to $\beta = [\beta_1, \beta_2]^{\mathsf{T}} = 1_2$ and the predictors generated as independent draws from a uniform distribution on $[-1, 2]$. The sample sizes were $n \in \{20, 80\}$ and $r = 10$. Code for reproducing the results is available at `https://github.com/koekvall/conf-crit`.

Figure 2 summarizes the results of a Monte Carlo experiment with 10,000 replications. Notably, the proposed confidence region has near-nominal estimated coverage probability for all considered settings; when $n = 80$, the nominal coverage probability 0.95 is within 2 times the Monte Carlo standard error for all considered λ. By contrast, the estimated coverage probabilities for the likelihood ratio and Wald confidence regions are substantially different from nominal for almost all considered settings. Moreover, their coverage is sometimes lower than nominal and sometimes higher than nominal. That is, for those methods, the quality of the chi-square distribution as a reference distribution depends on how close to the critical point the true parameter is. These results are consistent with the asymptotic theory which says the proposed test statistic is the only one of those considered whose asymptotic distribution is the same regardless of how a sequence of parameters approaches a critical point. We also note that in order to get an interesting comparison, we used the Wald test statistic standardized by expected Fisher information evaluated at the (unconstrained) maximum likelihood estimates; if one instead standardizes by expected information evaluated at the maximum likelihood estimates under the null hypothesis, performance is substantially worse than that reported here because that information matrix is singular with probability one.

Figure 3 examines the agreement between sample quantiles of the three considered test statistics and the theoretical quantiles of a chi-square distribution with three degrees of freedom. The comparison concerns two non-critical points, so classical asymptotic theory says all three test statistics have asymptotic chi-square distributions with three degrees of freedom. The first plot, corresponding to small but non-zero scale parameters, shows the approximation is poor near critical points for the likelihood ratio and Wald test statistics. For the larger scale parameters, agreement between sample and theoretical quantiles is decent for all three test statistics.

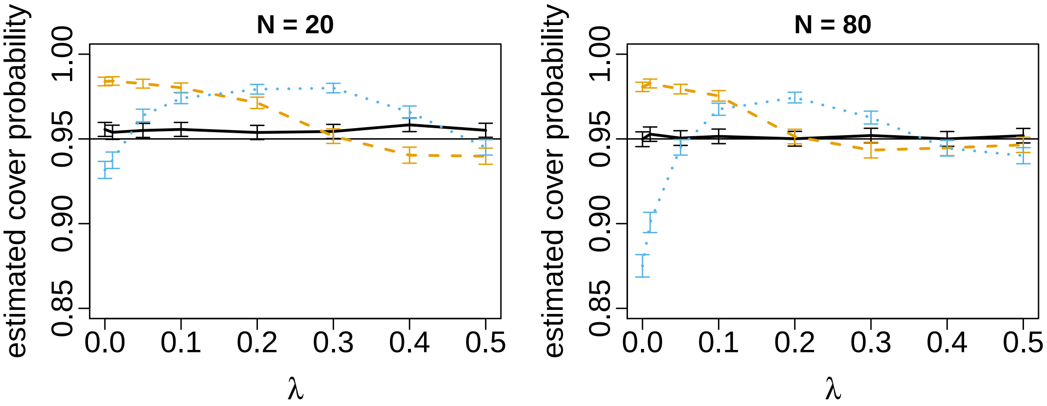

Figure 2: Monte Carlo estimates of coverage probabilities of confidence regions from inverting the modified score (solid), likelihood ratio (dashed), and Wald (dotted) test statistics. The straight horizontal line indicates the nominal 0.95 coverage probability and vertical bars denote ±2 times Monte Carlo standard errors.

For additional insight into how the test statistics, and hence the confidence regions, behave near critical points, we report in Figure 4 estimated rejection probabilities (size and power)

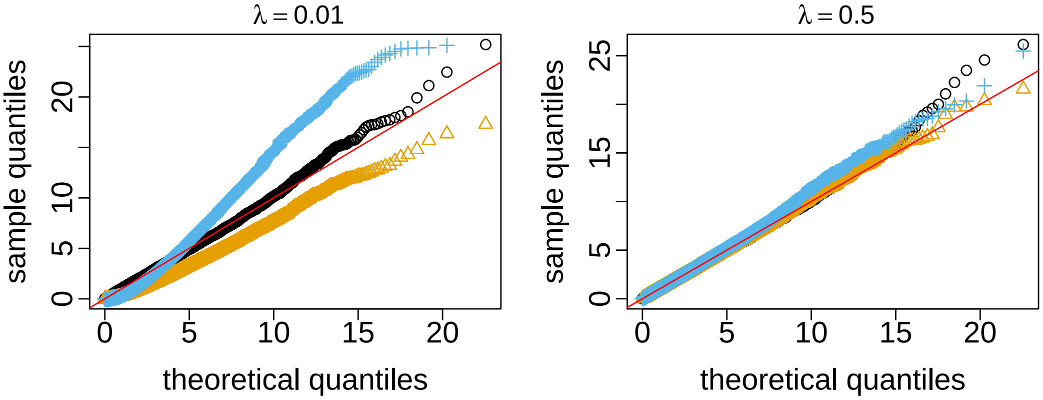

Figure 3: Quantile-quantile plots for modified score (circles), likelihood ratio (triangles), and Wald (plus signs) test statistics. The theoretical quantiles are from the chi-square distribution with 3 degrees of freedom and the sample quantiles from 10,000 Monte Carlo replications with $(n, r) = (80, 10)$.

of tests of the null hypothesis that $(\sigma, \lambda_1, \lambda_2) = (1, 0, 0)$. The data generating settings are the same as those used to generate the data in Figure 2. The power curves are not directly comparable because, as was also shown in Figure 2, the different tests have different sizes. Nevertheless, the power curves behave similarly as the true $\lambda_1 = \lambda_2$ moves away from the null hypothesis value. This indicates the differences in coverage observed in Figure 2 is not in general due to how large the different confidence regions are.

We considered several configurations in addition to those reported, including both larger and smaller values of n and λ, and the results were remarkably consistent. To compute the proposed test statistic, we used the `lmmstest` R package written by the first author. To fit the model and compute the likelihood ratio and Wald test statistics we used the `lme4` R package (Bates et al., 2015). There is also a Stata routine for calculation of the proposed confidence region for a single variance parameter in linear mixed models with a random intercept (Bottai and Orsini, 2004).

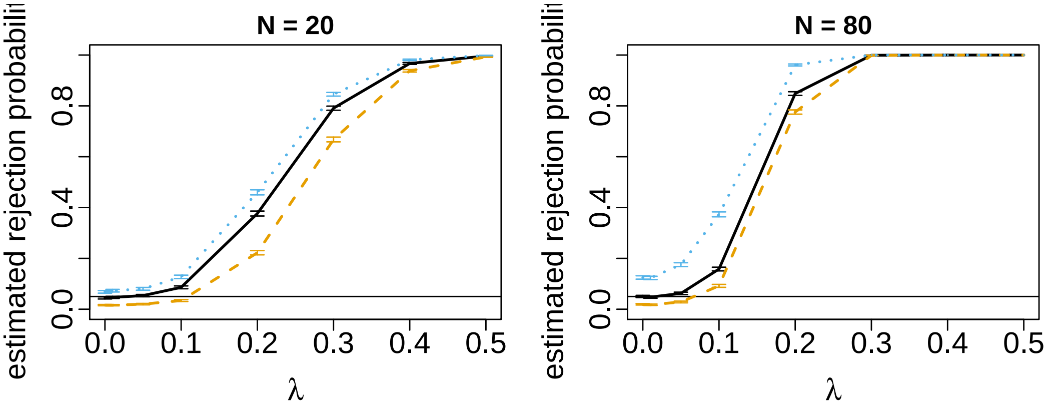

Figure 4: Monte Carlo estimates of rejection probabilities for testing the null hypothesis $(\sigma, \lambda_1, \lambda_2) = (1, 0, 0)$ using the modified score (solid), likelihood ratio (dashed), and Wald (dotted) test. The straight horizontal line indicates the size of the tests (0.05) and vertical bars denote ±2 times Monte Carlo standard errors.

5 Data example

We consider a dataset presented by Fitzmaurice et al. (2012), which is available at `https://content.sph.harvard.edu/fitzmaur/ala/` and contains a subset of the pulmonary function data collected in the Six Cities Study (Dockery et al., 1983). The data include a pulmonary measure called the forced expiratory volume in the first second (`FEV1`), height (`ht`), and age obtained from a randomly selected subset of the female participants living in Topeka, Kansas. The sample includes $n = 300$ girls and young women, with a minimum of one and a maximum of twelve observations over time. Age and height are believed to be associated with the ability to take in and force out air. To model these data, consider the linear mixed model

$$\texttt{FEV1}_{it} = \beta_1 + \beta_2\texttt{age}_{it} + \beta_3\texttt{ht}_{it} + \beta_4\texttt{age}_{i1} + \beta_5\texttt{ht}_{i1} + U_{i1} + U_{i2}\texttt{age}_{it} + E_{it}, \tag{10}$$

where $U_{i1} \sim \mathcal{N}(0, \lambda_1^2)$, $U_{i2} \sim \mathcal{N}(0, \lambda_2^2)$, and $E_{it} \sim \mathcal{N}(0, \sigma^2)$ are mutually independent for all individuals indexed by i and time points indexed by t. This is a model considered by Fitzmaurice et al. (2012), modified slightly to fit our setting.

The maximum likelihood estimate of β is $(-2.2, 0.078, 2.80, -0.040, -0.19)$, but we focus on the scale parameters whose estimates and confidence intervals are in Table 1. Notably, the maximum likelihood estimate of λ_1 is zero, indicating common confidence intervals may be

Parameter	Estimate	Mod. Score CI	Lik. Rat. CI	Wald CI
λ_1 (intercept)	0	$(0, 0.0482)$	$(0, 0.0604)$	$(0, 0.0772)$
λ_2 (age)	0.0201	$(0.0188, 0.0219)$	$(0.0183, 0.0222)$	$(0.0180, 0.0220)$
σ (error)	0.156	$(0.152, 0.162)$	$(0.151, 0.162)$	$(0.151, 0.162)$

Table 1: Maximum likelihood estimates and 95% confidence intervals (CI) for scale parameters in the linear mixed model (10).

unreliable. The proposed interval for λ_1 is substantially smaller than that based on the likelihood ratio. This is consistent with our simulations where the latter had greater than nominal empirical coverage of small scale parameters. One-dimensional Wald intervals using expected information are complicated to obtain because the information matrix at the maximum likelihood estimate is singular. Therefore, we present Wald intervals using observed information for the squared scale parameters, transformed to intervals for the scale parameters by taking square roots of the endpoints. The resulting interval for λ_1 is even wider than the likelihood ratio-based interval and hence the proposed interval is preferred.

The proposed interval for λ_2 is substantially narrower than the other two, and its left endpoint is further from zero. Thus, the proposed interval leads to not only reliable but more precise inference. The intervals for σ only differ in the third significance digit. This is consistent with both theory and simulations since the estimate of σ is further from zero than those of λ_1 and λ_2, and hence the different test statistics are expected to behave similarly.

Figure 5 shows plots of the proposed test-statistics for a range of λ_1 and λ_2. The values of λ_1 and λ_2 such that the graph of the corresponding test-statistic is below the critical value 3.84, the 0.95th quantile of the chi-square distribution with one degree of freedom, gives the confidence regions in Table 1. The graphs indicate the test statistics are convex in λ_1 and λ_2, respectively. The third plot in Figure 5 depicts confidence regions for $\lambda = (\lambda_1, \lambda_2)$ using the proposed procedure. The regions can be used to assess which values of λ are supported by the data and we may, for example, reject the joint null hypothesis that $\lambda_1 = \lambda_2 = 0$ at conventional levels of significance.

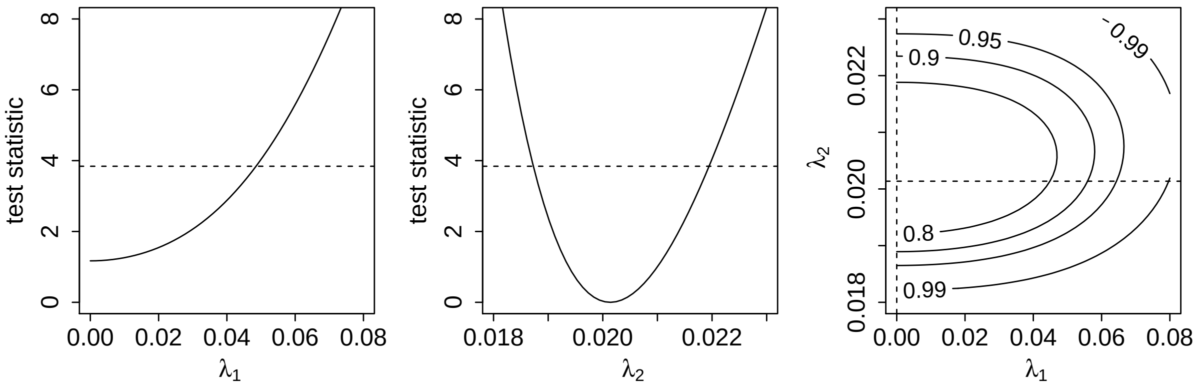

Figure 5: Separate test statistics (first and second plot) and joint 80 – 99 % confidence regions for λ (third plot). The dashed lines in the first two plots mark the 0.95th quantile of a chi-square distribution with one degree of freedom. The dashed lines in the third plot mark maximum likelihood estimates.

6 Final remarks

Linking the boundary problem to the singular information problem allows a deeper understanding of the behavior of the likelihood function in shrinking neighborhoods of the boundary of a parameter set. Perhaps more importantly, it permits the construction of confidence regions that have valid asymptotic uniform coverage probability.

The advantages of using the proposed modified score test in constructing confidence regions are many-fold: the proposed procedure does not require a point estimate, which can be troublesome when the parameter is at or near the boundary; it does not rely on simulation algorithms, which typically need to be programmed for the specific problem at hand; it can be applied to a broad variety of models, of which the linear mixed-effects model is but a special case; it allows inference on scale parameters when the random effects follow an asymmetric distribution, which gives insight about the sign and the magnitude of the skewness. In addition, to the best of our knowledge, the asymptotic behavior of the Wald test and likelihood ratio test for a scale parameter has not been described for the general setting in which the information matrix has any rank less than full.

Avenues for future research include theory for confidence regions for a sub-vector of the parameter when the Fisher information is not block-diagonal. Then, loosely speaking, inference using the profile likelihood is no longer equivalent to inference using the full likelihood with plug-in estimators, and hence new theory is needed. Efficient software implementations for popular mixed models are also needed.

References

Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families.* Number 3 in Institute of Mathematical Statistics monographs. Cambridge University Press, Cambridge.

Baey, C., Cournède, P.-H., and Kuhn, E. (2019). Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 135.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).

Billingsley, P. (1995). *Probability and Measure.* Wiley series in probability and mathematical statistics. Wiley, New York, 3rd ed edition.

Billingsley, P. (1999). *Convergence of Probability Measures.* Wiley series in probability and statistics. Probability and statistics section. Wiley, New York, second edition.

Bottai, M. (2003). Confidence regions when the Fisher information is zero. *Biometrika*, 90(1).

Bottai, M. and Orsini, N. (2004). Confidence intervals for the variance component of random-effects linear models. *The Stata Journal: Promoting communications on statistics and Stata*, 4(4).

Chen, S. T., Xiao, L., and Staicu, A.-M. (2019). An approximate restricted likelihood ratio test for variance components in generalized linear mixed models. *arXiv:1906.03320 [stat].*

Chesher, A. (1984). Testing for neglected heterogeneity. *Econometrica*, 52(4).

Cox, D. R. and Hinkley, D. V. (2000). *Theoretical Statistics.* Chapman & Hall/CRC, Boca Raton.

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 (Statistical Methodology)*, 66(1).

Dockery, D., Berkey, C., Ware, J., Speizer, F., and Ferris Jr, B. (1983). Distribution of forced vital capacity and forced expiratory volume in one second in children 6 to 11 years of age. *American Review of Respiratory Disease*, 128(3).

Drikvandi, R., Verbeke, G., Khodadadi, A., and Partovi Nia, V. (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics*, 14(1).

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied Longitudinal Analysis.* John Wiley & Sons, Hoboken, NJ.

Fitzmaurice, G. M., Lipsitz, S. R., and Ibrahim, J. G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics*, 63(3).

Fristedt, B. E. and Gray, L. F. (2013). *A Modern Approach to Probability Theory.* Birkhäuser Boston, Boston, MA.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *Annals of Statistics*, 22(4).

Giampaoli, V. and Singer, J. M. (2009). Likelihood ratio tests for variance components in linear mixed models. *Journal of Statistical Planning and Inference*, 139(4).

Greven, S., Crainiceanu, C. M., Küchenhoff, H., and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17(4).

Hall, D. B. and Praestgaard, J. T. (2001). Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika*, 88(3).

Lee, L.-F. and Chesher, A. (1986). Specification testing when score test statistics are identically zero. *Journal of Econometrics*, 31(2).

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2).

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, linear, and mixed models*. John Wiley & Sons, Hoboken, NJ.

Qu, L., Guennel, T., and Marshall, S. L. (2013). Linear score tests for variance components in linear mixed models and applications to genetic association studies: linear score tests for variance components. *Biometrics*, 69(4).

Rotnitzky, A., Cox, D. R., Bottai, M., and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, 6(2).

Saville, B. R. and Herring, A. H. (2009). Testing random effects in the linear mixed model using approximate Bayes factors. *Biometrics*, 65(2).

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J, 2nd ed edition.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398).

Sinha, S. K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *Canadian Journal of Statistics*, 37(2).

Stern, S. E. and Welsh, A. H. (2000). Likelihood inference for small variance components. *Canadian Journal of Statistics*, 28(3).

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4).

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, 59(2).

Wiencierz, A., Greven, S., and Küchenhoff, H. (2011). Restricted likelihood ratio testing in linear mixed models with general error covariance structure. *Electronic Journal of Statistics*, 5(0).

Wood, S. N. (2013). A simple test for random effects in regression models. *Biometrika*, 100(4).

Zhu, H. and Zhang, H. (2006). Generalized score test of homogeneity for mixed effects models. *The Annals of Statistics*, 34(3).

A Technical details

A.1 Proofs of results in main text

Proof of Lemma 2.2. The assumptions of the lemma imply $T_n(\cdot;\cdot)$ is continuous on $\{\theta : \mathcal{I}(\theta) > 0\} \times \mathcal{Y}^n$. They also say we may, for any critical θ and $(y_1, \dots, y_n) \in \mathcal{Y}^n$, unambiguously define $T_n(\theta; y_1, \dots, y_n) = \lim_{m\to\infty} T_n(\theta_m; y_1, \dots, y_n)$, where $\{\theta_m\}$ is any sequence of non-critical points tending to θ; Assumption 5 says at least one such sequence exists. To verify this extension is continuous on $\Theta \times \mathcal{Y}^n$, let instead $\{\theta_m\} \in \Theta$ be an arbitrary sequence, possibly including critical points, tending to θ. Let also $\{(y_{m1}, \dots, y_{mn})\} \in \mathcal{Y}^n$ be an arbitrary sequence tending to $(y_1, \dots, y_n)$. By the assumptions of the lemma, we can find, for every fixed m, a non-critical $\tilde{\theta}_m$ such that

$$|T_n(\theta_m; y_{m1}, \dots, y_{mn}) - T_n(\tilde{\theta}_m; y_{m1}, \dots, y_{mn})| \leq 1/m \text{ and } \|\theta_m - \tilde{\theta}_m\| \leq 1/m.$$

Thus, by the triangle inequality,

$$|T_n(\theta_m; y_{m1}, \dots, y_{mn}) - T_n(\theta; y_1, \dots, y_n)| \leq 1/m + |T_n(\tilde{\theta}_m; y_{m1}, \dots, y_{mn}) - T_n(\theta; y_1, \dots, y_n)|,$$

which tends to zero by the assumptions of the lemma since $\{\tilde{\theta}_m\}$ is a sequence of non-critical points tending to θ. □

Proof of Lemma 2.5. Suppose for contradiction there is a compact $C \subseteq \Theta$ and an $\epsilon > 0$ such that, for infinitely many n, $\sup_{\theta\in C} |\mathsf{P}_\theta \{\theta \in \mathcal{R}_n(\alpha)\} - (1 - \alpha)| > \epsilon$. Let N be the set of such n and pick, for every $n \in N$, a $\theta_n \in C$ such that $|\mathsf{P}_{\theta_n} \{\theta_n \in \mathcal{R}_n(\alpha)\} - (1 - \alpha)| > \epsilon$ Because C

is compact, it is bounded and hence $\{\theta_n : n \in N\}$ is a bounded sequence. Thus, it contains a convergent subsequence, say $\{\theta_n : n \in N_1\}$, $N_1 \subseteq N$. But by assumption, along this subsequence, $T_n(\theta_n; Y_{n1}, \ldots, Y_{nn}) \rightsquigarrow \chi^2_d$; in particular, since χ^2_d has a continuous cumulative distribution function, $\mathsf{P}_{\theta_n}\{\theta_n \in \mathcal{R}_n(\alpha)\} = \mathsf{P}\{T_n(\theta_n; Y_{n1}, \ldots, Y_{nn}) \leq q_{d,1-\alpha}\} \to 1-\alpha$, which is the desired contradiction. We have thus proven, for every compact $C \subseteq \Theta$ and $\epsilon > 0$, $\sup_{\theta\in C} |\mathsf{P}_\theta\{\theta \in \mathcal{R}_n(\alpha)\} - (1-\alpha)| \leq \epsilon$ for all but at most finitely many n. Thus, for any compact $C \subseteq \Theta$ and $\epsilon > 0$, $1-\alpha-\epsilon \leq \liminf_{n\to\infty} \inf_{\theta\in C} \mathsf{P}_\theta\{\theta \in \mathcal{R}_n(\alpha)\} \leq \limsup_{n\to\infty} \inf_{\theta\in C} \mathsf{P}_\theta\{\theta \in \mathcal{R}_n(\alpha)\} 1-\alpha+\epsilon$, and sending $\epsilon \to 0$ finishes the proof. □

Proof of Theorem 2.4. By Lemma 2.6, it suffices to consider an arbitrary sequence $\{\theta_n\}$ of non-critical points tending to a $\theta \in \Theta$. For any $t \in \mathbb{R}^d$, let $U_{nt} = n^{-1/2}\sum_{i=1}^n t^\mathsf{T} A_n V^\mathsf{T} s(\theta_n, Y_{ni}) = t^\mathsf{T} U_n$, where $V = [v_1, \ldots, v_d] = [v_{\theta 1}, \ldots, v_{\theta d}] \in \mathbb{R}^{d\times d}$ a matrix whose columns satisfy Assumption 3 at θ and $A_n = \operatorname{diag}(a_{n1}, \ldots, a_{nd})$ is a scaling matrix defined by the $\{a_{nj}\}$ given by Lemma 2.3 (with the index $m = n$), and $Y_{n1}, \ldots, Y_{nn}$ are independent with common density f_{θ_n}. We have $\mathbb{E}(U_{nt}) = 0$ since the score function has mean zero by the first Bartlett identity, applicable owing to Assumption 2, and by the convergence of covariance matrices established in the proof of Theorem 2.1 and Assumption 4,

$$\operatorname{var}(U_{nt}) = \operatorname{var}\{t^\mathsf{T} A_n V^\mathsf{T} s(\theta_n, Y_{n1})\} \to t^\mathsf{T} \operatorname{cov}_\theta\{\tilde{s}(\theta, Y)\} t > 0.$$

Moreover, the jth element of $A_n V^\mathsf{T} s(\theta_n, Y_{n1})$ is equal to $v_j^\mathsf{T} s(\theta_n, Y_{n1})$ if θ_j is critical and $\partial^k v_j^\mathsf{T} s(\tilde{\theta}^{(j)}, Y_{n1})/\partial\theta_j^k$ otherwise, where $\tilde{\theta}_n^{(j)}$ a point between θ_n and θ selected as in the proof of Lemma 2.3. Thus, by definition of spectral norm, $\|A_n V^\mathsf{T} s(\theta_n, Y_{n1})\| \|t\| \geq |t^\mathsf{T} A_n V^\mathsf{T} s(\theta_n, Y_{n1})|$, and therefore the last right-hand side has $2+\delta$ bounded moments for all large enough n by Assumption 2. Thus, Lyapunov's condition (Billingsley, 1995, 27.16) is satisfied and, by Slutsky's theorem (Billingsley, 1999, Theorem 3.1), $U_{nt} \rightsquigarrow \mathcal{N}(0, t^\mathsf{T} \operatorname{cov}_\theta\{\tilde{s}(\theta, Y)\} t)$. Thus, by the Cramér–Wold theorem (Billingsley, 1995, Theorem 29.4) $U_n \rightsquigarrow \mathcal{N}[0, \operatorname{cov}_\theta\{\tilde{s}(\theta, Y)\}]$. Now the result follows from the continuous mapping theorem (Billingsley, 1999, Theorem 2.7) since $T_n(\theta_n; Y_{n1}, \ldots, Y_{nn}) = U_n^\mathsf{T}\{A_n V^\mathsf{T} \mathcal{I}(\theta_n) V A_n\}^{-1} U_n$. □

Proof of Proposition 3.3. Assumption 3 is immediate from the assumption that, for any θ, the null space is spanned by $\{e_j, j : \lambda_j = 0\}$. Then Assumption 5 follows: if θ is a critical point, then the rank of $\mathcal{I}(\theta)$ is the number of $\lambda_j = 0$. Take the sequence given by (ii) but fix $\lambda_{mj} \neq 0$ for all m and j such that $\lambda_j \neq 0$. The resulting sequence is a sequence of non-critical points tending to θ. Finally, consider an arbitrary θ and suppose there are k critical elements, without loss of generality the first k: $\lambda_1 = \cdots = \lambda_k = 0$. Let $V = [e_1, \dots, e_k, v_{\theta(k+1)}, \dots, v_{\theta d}]$ have columns that are orthonormal eigenvectors of $\mathcal{I}(\theta)$. A straightforward calculation using that $v_{\theta j}$, $j \geq k+1$, has zeros in the first k elements by orthogonality shows $\tilde{s}(\theta, Y) = V^\mathsf{T} \bar{s}(\theta, Y)$, and from this it is clear that $\operatorname{cov}_\theta\{\bar{s}(\theta, Y)\} > 0$ is equivalent to $\operatorname{cov}_\theta\{\tilde{s}(\theta, Y)\} > 0$. □

Proof of Lemma 3.7. Differentiating the log-likelihood with respect to β shows the last d_2 elements of $s(\theta, y)$ have the familiar form $s_\beta(\theta, y_i, X_i) = X_i^\mathsf{T} \Sigma^{-1}(y_i - X_i\beta)$. Differentiating this with respect to λ and taking expectations shows $\mathcal{I}(\theta)$ is block-diagonal. The trailing block is $\mathcal{I}_\beta(\theta) = \operatorname{cov}_\theta\{s_\beta(\theta, Y_i, X_i)\} = \mathbb{E}(X^\mathsf{T}\Sigma^{-1}X)$, which is positive definite for all θ since $\underline{\mathrm{e}}(\Sigma) \geq \sigma^2 > 0$ and $\underline{\mathrm{e}}\{\mathbb{E}(X^\mathsf{T}X)\} > 0$ by assumption; the result follows. □

Proof of Lemma 3.8. For $j = 1, \dots, d_1$, let

$$\xi_j(\theta, y, X) = \operatorname{tr}\left\{\Sigma^{-1}H_j - \Sigma^{-1}(y - X\beta)(y - X\beta)^\mathsf{T}\Sigma^{-1}H_j\right\}$$

so that $s_j(\theta, y, X) = \lambda_j \xi_j(\theta, y, X)$. Then claim to be proved is equivalent to, for any $v \in \mathbb{R}^{d_1}$,

$$\operatorname{var}_\theta\{v^\mathsf{T}\xi(\theta, Y, X)\} \geq 2\,\bar{\mathrm{e}}(\Sigma)^{-2}\,\underline{\mathrm{e}}(Z^\mathsf{T}Z)\min_j \|Z^j\| \max_j (v_j)^2,$$

where $\xi = [\xi_1, \dots, \xi_d]^\mathsf{T}$. With $G = \sum_{j=1}^d v_j H_j \in \mathbb{R}^{r\times r}$, we have

$$v^\mathsf{T}\xi(\theta, Y, X) = \operatorname{tr}\left[\left\{\Sigma^{-1} - \Sigma^{-1}(Y - X\beta)(Y - X\beta)^\mathsf{T}\Sigma^{-1}\right\}G\right].$$

Thus, applying the well-known expression for the variance of a quadratic form in multivariate normal vectors (Seber and Lee, 2003, Theorem 1.6),

$$\begin{aligned}
\operatorname{var}_\theta\left\{v^\mathsf{T}\xi(\theta, Y, X) \mid X\right\} &= \operatorname{var}_\theta\left[\operatorname{tr}\left\{\Sigma^{-1}(Y - X\beta)(Y - X\beta)^\mathsf{T}\Sigma^{-1}G\right\} \mid X\right] \\
&= \operatorname{var}_\theta\left[(Y - X\beta)^\mathsf{T}\Sigma^{-1}G\Sigma^{-1}(Y - X\beta) \mid X\right] \\
&= 2\operatorname{tr}[(\Sigma^{-1/2}G\Sigma^{-1/2})^2].
\end{aligned}$$

Now observe that since $\Sigma^{-1/2}G\Sigma^{-1/2}$ is symmetric, its eigenvalues are real, and hence the eigenvalues of $(\Sigma^{-1/2}G\Sigma^{-1/2})^2$ are non-negative as the squares of real numbers. Thus, the trace upper bounds the maximum eigenvalue and we get

$$\begin{aligned}\mathbb{E}_\theta\left[\operatorname{var}_\theta\left\{v^\mathsf{T}\xi(\theta,Y,X)\mid X\right\}\right] &= \operatorname{var}_\theta\left\{v^\mathsf{T}\xi(\theta,Y,X)\mid X\right\}\\ &\geq 2\|(\Sigma^{-1/2}G\Sigma^{-1/2})^2\|\\ &= 2\|\Sigma^{-1/2}G\Sigma^{-1}G\Sigma^{-1/2}\|\\ &\geq 2\,\underline{\mathrm{e}}(\Sigma^{-1})\|\Sigma^{-1/2}G^2\Sigma^{-1/2}\|\\ &\geq 2\,\underline{\mathrm{e}}(\Sigma^{-1})^2\,\bar{\mathrm{e}}(G^2).\end{aligned}$$

Now write

$$G=\sum_{j=1}^{d}\sum_{k\in[j]}v_jZ^k(Z^k)^\mathsf{T}=\sum_{k=1}^{r}v_{j(k)}Z^k(Z^k)^\mathsf{T}=Z\tilde{V}Z^\mathsf{T},$$

where $\tilde{V}$ is diagonal with the elements of v on the diagonal, ordered so that the last equality holds. Then

$$\|G^2\|=\|Z\tilde{V}Z^\mathsf{T}Z\tilde{V}Z^\mathsf{T}\|\geq\underline{\mathrm{e}}(Z^\mathsf{T}Z)\|Z\tilde{V}^2Z^\mathsf{T}\|.$$

The last norm is $\|Z\tilde{V}^2Z^\mathsf{T}\|=\bar{\mathrm{e}}(Z\tilde{V}^2Z^\mathsf{T})=\max_{\|b\|=1}b^\mathsf{T}Z\tilde{V}^2Z^\mathsf{T}b$ which by considering $b=Z^j/\|Z^j\|$ is upper bounded by, for every $j=\dots,q$,

$$\left(\frac{Z^j}{\|Z^j\|}\right)^\mathsf{T}\sum_{k=1}^{r}v_{j(k)}^2Z^k(Z^k)^\mathsf{T}\left(\frac{Z^j}{\|Z^j\|}\right)\geq\max_k v_{j(k)}^2\min_k\|Z^k\|^2$$

We have shown

$$\mathbb{E}_\theta[\operatorname{var}_\theta\{v^\mathsf{T}\xi(\theta,Y,X)\mid X\}]\geq 2\,\underline{\mathrm{e}}(\Sigma^{-1})^2\,\underline{\mathrm{e}}(Z^\mathsf{T}Z)\max_k v_{j(k)}^2\min_k\|Z^k\|^2,$$

and the proof is completed by observing that $\mathbb{E}_\theta\{v^\mathsf{T}\xi(\theta,Y,X)\mid X\}=0$ so that, by the law of total variance, $\operatorname{var}_\theta\{v^\mathsf{T}\xi(\theta,Y,X)\}=\mathbb{E}_\theta[\operatorname{var}_\theta\{v^\mathsf{T}\xi(\theta,Y,X)\mid X\}]$. □

Proof of Theorem 3.9. Assumption 1 is satisfied since Y has support on $\mathbb{R}^r$ for all θ, with γ being Lebesgue measure. We verify Assumption 2 with $\delta>0$ and $k=2$. For $j,k\leq d_1$,

$$\begin{aligned}\frac{\partial s_j(\theta,y)}{\partial\lambda_k} &= \mathbb{I}(j=k)\frac{s_j(\theta,y)}{\lambda_j}\\ &\quad-\lambda_j\operatorname{tr}\left\{\Sigma^{-1}H_k\Sigma^{-1}H_j\right\}\\ &\quad+\lambda_j\operatorname{tr}\left\{\Sigma^{-1}H_k\Sigma^{-1}(y-X\beta)(y-X\beta)^\mathsf{T}\Sigma^{-1}H_j\right\}\\ &\quad+\lambda_j\operatorname{tr}\left\{\Sigma^{-1}(y-X\beta)(y-X\beta)^\mathsf{T}\Sigma^{-1}H_k\Sigma^{-1}H_j\right\}.\end{aligned}$$

By picking B small enough, we may assume $\sigma^2 \geq \epsilon > 0$ on B. Thus, $\|\Sigma^{-1}\| \leq \epsilon^{-1}$ on B. By decreasing ϵ if needed, we may also assume $\lambda_j \leq \epsilon^{-1}$ on B for all j and $\|\beta\| \leq \epsilon^{-1}$. Thus, using sub-multiplicativity of the spectral norm and that the trace is upper bounded by r times the spectral norm, on B,

$$\begin{aligned}\left|\frac{\partial s_j(\theta, y)}{\partial \lambda_k}\right| &\leq r\left\{\epsilon^{-1}\|H_j\| + \epsilon^{-2}\|H_j\|(2\|y\|^2 + 2\|X\|^2\epsilon^{-2})\right\} \\ &\quad + \epsilon^{-1} r\left\{\|H_j\|\|H_k\|\epsilon^{-2}\right\} \\ &\quad + 2\epsilon^{-1} r\left\{\epsilon^{-3}\|H_k\|\|H_j\|(2\|y\|^2 + 2\|X\|^2\epsilon^{-2})\right\},\end{aligned}$$

Thus, the expectation whose supremum is to be bounded is less than

$$\mathbb{E}_\theta(c_1 + c_2\|Y\|^{4+\delta} + c_3\|X\|^{4+\delta})$$

for some finite and positive c_1, c_2, c_3 who can depend on Z. This expectation is bounded for $\theta \in B \cap \Theta$ since the mean of Y and the eigenvalues of its covariance matrix are both bounded for such θ. The calculation for $l = 0$ in Assumption 2 is very similar so we omit it. To verify the remaining assumptions, we use Proposition 3.3. Its condition (i) clearly holds, (ii) holds by Lemmas 3.7 and 3.8, and that (iii) holds is essentially verified in the proof of Lemma 3.8. Specifically, the vector there denoted $\xi(\theta, Y)$, with elements $\xi_j(\theta, Y) = \operatorname{tr}\left\{\Sigma^{-1}H_j - \Sigma^{-1}(y - X\beta)(y - X\beta)^\mathsf{T}\Sigma^{-1}H_j\right\}$, is positive definite if and only if $\bar{s}(\theta, Y)$ is. To see this, note that ξ and $\bar{s}$ are the same up to elementwise multiplication by a scalar that is equal one if $\lambda_j = 0$ and is equal to λ_j otherwise. □

Proof of Theorem 3.10. The claim about $\mathcal{R}_n^\lambda(\alpha)$ is almost immediate from Theorem 3.9 and the fact that $\mathcal{I}(\theta)$ is block diagonal so we omit the proof. To prove the second claim, observe that $T_n^\lambda(\lambda; \beta)$ can be written as

$$T_n^\lambda(\lambda; \beta) = n^{-1}\left\{\sum_{i=1}^n \xi(\theta, Y_i, X_i)^\mathsf{T}\right\}\operatorname{cov}_\theta\{\xi(\theta, Y)\}^{-1}\left\{\sum_{i=1}^n \xi(\theta, Y_i, X_i)\right\},$$

where $\xi_j(\theta, Y_i, X_i) = \operatorname{tr}\left\{\Sigma^{-1}H_j - \Sigma^{-1}(Y_i - X_i\beta)(Y_i - X_i\beta)^\mathsf{T}\Sigma^{-1}H_j\right\}$. The inverse covariance matrix is continuous in θ by the arguments in the proof of Theorem 2.1. Thus, it suffices (Billingsley, 1999, Theorems 2.7 and 3.1) to show, with $(Y_{ni}, X_{ni}) \sim f_{\theta_n}$,

$$\left\|n^{-1/2}\sum_{i=1}^n \xi(\lambda_n, \beta_n, Y_{ni}, X_{ni}) - n^{-1/2}\sum_{i=1}^n \xi(\lambda_n, \hat{\beta}_n, Y_{ni}, X_{ni})\right\| = o_\mathsf{P}(1).$$

We show the equivalent result that every element of the vector in the norm is $o_{\mathsf{P}}(1)$. The jth element is

$$n^{-1/2}\sum_{i=1}^{n}\left\{(Y_{ni}-X_{ni}\beta_n)^{\mathsf{T}}\Omega_{nj}(Y_{ni}-X_{ni}\beta_n)-(Y_{ni}-X_{ni}\hat{\beta}_n)^{\mathsf{T}}\Omega_{nj}(Y_{ni}-X_{ni}\hat{\beta}_n)\right\},$$

where $\Omega_{nj}=\Sigma_n^{-1}H_j\Sigma_n^{-1}$. Let $\varepsilon_{ni}=Y_{ni}-X_{ni}\beta_n\sim\mathcal{N}(0,\Sigma_n)$ to get that the last display is equal to

$$n^{-1/2}\sum_{i=1}^{n}\left[\varepsilon_{ni}^{\mathsf{T}}\Omega_{nj}\varepsilon_{ni}-\{\varepsilon_{ni}+X_{ni}(\beta_n-\hat{\beta}_n)\}^{\mathsf{T}}\Omega_{nj}\{\varepsilon_{ni}+X_{ni}(\beta_n-\hat{\beta}_n)\}\right],$$

which in turn is equal to

$$-n^{-1/2}2(\beta_n-\hat{\beta}_n)^{\mathsf{T}}\sum_{i=1}^{n}X_{ni}^{\mathsf{T}}\Omega_{nj}\varepsilon_{ni}-n^{-1/2}(\beta_n-\hat{\beta}_n)^{\mathsf{T}}\left(\sum_{i=1}^{n}X_{ni}^{\mathsf{T}}\Omega_{nj}X_{ni}\right)(\beta_n-\hat{\beta}_n).$$

Thus, since $\|\beta_n-\hat{\beta}_n\|=O_{\mathsf{P}}(1/\sqrt{n})$ it suffices to show that

$$\left\|n^{-1}\sum_{i=1}^{n}X_{ni}^{\mathsf{T}}\Omega_{nj}\varepsilon_{ni}\right\|=o_{\mathsf{P}}(1)\ \text{ and }\ \left\|n^{-1}\sum_{i=1}^{n}X_{ni}^{\mathsf{T}}\Omega_{nj}X_{ni}\right\|=O_{\mathsf{P}}(1).$$

The second holds since, for any $t\in\mathbb{R}^{d_2}$ of unit length, $t^{\mathsf{T}}X_{ni}^{\mathsf{T}}\Omega_{nj}X_{ni}t\leq\|\Omega_{nj}\|t^{\mathsf{T}}X_{ni}^{\mathsf{T}}X_{ni}t$, $\|\Omega_{nj}\|\leq\|H_j\|\|\Sigma^{-1}\|\leq\|H_j\|\sigma^{-4}$, and $n^{-1}\sum_{i=1}^{n}t^{\mathsf{T}}X_{ni}^{\mathsf{T}}X_{ni}t\to t^{\mathsf{T}}\mathbb{E}(X^{\mathsf{T}}X)t\leq\mathbb{E}(\|X\|^2)<\infty$ by the law of large numbers and, for the last inequality, assumption. To show the first condition in the last display, condition on $\{X_1,\dots,X_n\}$, apply Chebyshev's inequality, and take expectations to get, for any $s\geq 0$ and $t\in\mathbb{R}^r$,

$$\begin{aligned}\mathsf{P}\left\{\left|t^{\mathsf{T}}n^{-1}\sum_{i=1}^{n}X_{ni}^{\mathsf{T}}\Omega_{nj}\varepsilon_{ni}\right|\geq s\right\}&\leq\frac{1}{s^2n^2}\mathbb{E}\left(\sum_{i=1}^{n}t^{\mathsf{T}}X_{ni}^{\mathsf{T}}\Omega_{nj}\Sigma_n\Omega_{nj}X_{ni}t\right)\\&\leq\frac{1}{s^2n}\|H_j\|^2\sigma^{-6}\|t\|^2\mathbb{E}(\|X\|^2),.\end{aligned}$$

Taking the supremum over t with $\|t\|=1$ and sending $n\to\infty$ finishes the proof. □

A.2 Additional and ancillary results

Lemma A.1. *For any $n\in\{1,2,\dots\}$ and sequence $\{\theta_m\}\in\Theta$ tending to some $\theta\in\Theta$, with $Y_{m1},\dots,Y_{mn}$ independent with density f_{θ_m} and $Y_1,\dots,Y_n$ independent with density f_θ, as $m\to\infty$ with n fixed:*

$$T_n(\theta_m;Y_{m1},\dots,Y_{mn})\rightsquigarrow T_n(\theta;Y_1,\dots,Y_n).$$

Proof. Since $f_{\theta_m} \to f_\theta$ pointwise by Assumption 2, $(Y_{m1}, \ldots, Y_{mn}) \to (Y_1, \ldots, Y_n)$ in total variation by Scheffe's lemma, and hence also in distribution. Thus, by Slutsky's theorem,

$$(\theta_m, Y_{m1}, \ldots, Y_{mn}) \rightsquigarrow (\theta, Y_1, \ldots, Y_n).$$

The result follows now follows from the continuous mapping theorem and Theorem 2.1. □

Theorem A.2. *Let f_θ be as defined in Example 1 with known $\psi = 0$, $r = 1$, and define the the score test standardized by observed information*

$$T_n^O(\theta, y_1, \ldots, y_n) = \left\{ \sum_{i=1}^n -\partial^2 \log f_\theta(y_i)/\partial\theta^2 \right\}^{-1} \left\{ \sum_{i=1}^n \partial \log f_\theta(y_i)/\partial\theta \right\}^2.$$

Then with $Y_{n1}, \ldots, Y_{nn}$ independent with density f_{θ_n}, and $Z \sim \mathcal{N}(0,1)$, as $n \to \infty$,

$$T_n^O(\theta_n; Y_{n1}, \ldots, Y_{nn}) \rightsquigarrow \begin{cases} Z^2 & \text{if } n^{1/4}|\theta_n| \to \infty \\ \frac{2a^2Z^2}{2a^2 - \sqrt{2}Z} & \text{if } \theta_n = an^{-1/4}, \quad a \in \mathbb{R} \\ 0 & \text{if } \theta_n = o(n^{-1/4}) \end{cases}$$

Proof. Recall $s(\theta, y) = \theta\{-1 + y^2/(1+\theta^2)\}/(1+\theta^2)$. Some algebra gives $\partial s(\theta, y)/\partial\theta = (\theta^4 - 3\theta^2 y^2 + y^2 - 1)/(1+\theta^2)^3$ and hence

$$T_n^O(\theta) = \theta^2(1+\theta^2)\frac{\left[\sum_{i=1}^n \{y_i/(1+\theta^2) - 1\}\right]^2}{\sum_{j=1}^n \{1 - y_i^2 + 3\theta^2 y_i^2 - \theta^4\}}$$

Let $x_n = \sum_{i=1}^n \{y_i^2/(1+\theta^2) - 1\}$, or $\sum_{i=1}^n y_i^2 = (1+\theta^2)(x_n + n)$, to get

$$\begin{aligned} T_n^O(\theta) &= \theta^2(1+\theta^2)\frac{x_n^2}{n(1+\theta^2)(1-\theta^2) - (1-3\theta^2)(x_n+n)(1+\theta^2)} \\ &= \frac{x_n^2/n}{2 - (1-3\theta^2)(\theta^2 n)^{-1} x_n} \end{aligned}$$

Observe that $X_n \sim (\chi_n^2 - n)$ regardless of θ, where X_n is defined as x_n but with Y_i in place of y_i. Thus, $n^{-1/2}X_n \rightsquigarrow \sqrt{2}Z$ by the central limit theorem. Thus, if $\theta_n^2 n = a^2\sqrt{n}$, or $\theta_n = an^{-1/4}$, then $T_n^O(\theta_n) \rightsquigarrow 2a^2Z^2/(2a^2 - \sqrt{2}Z)$ by Slutsky and mapping theorems. The other cases now follow by routine arguments. □

B Additional example details

Example 1. Recall, the $Y_i = [Y_1, \dots, Y_r]^\mathsf{T}$, $i = 1, \dots, n$, are independent and multivariate normally distributed with mean 0 and common covariance matrix $\Sigma(\theta) = \theta^2 1_r 1_r^\mathsf{T} + I_r$. Thus, the log-likelihood is for one observation is

$$\log f_\theta(y_i) = -\frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2} y_i^\mathsf{T} \Sigma(\theta)^{-1} y_i.$$

Since $1_r 1_r^\mathsf{T}$ has eigenvalues r and 0, $\Sigma(\theta)$ has eigenvalues $1 + r\theta^2$ and 1, the latter with multiplicity $r - 1$. Thus, $|\Sigma(\theta)| = (1 + r\theta^2)1^{r-1} = 1 + r\theta^2$. Applying the Sherman–Morrison formula to $\Sigma(\theta)^{-1}$ gives$(I_r + \theta^2 1_r 1_r^\mathsf{T})^{-1} = I_r - \theta^2 1_r 1_r^\mathsf{T}(1 + r\theta^2)^{-1}$, and hence $2\log f_\theta(y_i) = -\log(1 + r\theta^2) - y_i^\mathsf{T} y_i + (y_i^\mathsf{T} 1_r)^2 \theta^2 (1 + r\theta^2)^{-1}$. Differentiating $\log f_\theta(y_i)$ with respect to θ gives $s(\theta, y) = -r\theta(1 + r\theta^2)^{-1} + \theta(y_i^\mathsf{T} 1_r)^2(1 + r\theta^2)^{-2}$. Differentiating again we get $h(\theta, y) := -(r - r^2\theta^2)\{(1 + r\theta^2)^2\}^{-1} + (y_i^\mathsf{T} 1_r)^2(1 - 3r\theta^2)(1 + r\theta^2)^{-3}$. Thus, using that $\mathbb{E}_\theta\{(Y_i^\mathsf{T} 1_r)^2\} = \mathrm{var}_\theta(Y_i^\mathsf{T} 1_r) = 1_r^\mathsf{T} \Sigma(\theta) 1_r = 1_r^\mathsf{T} 1_r (1 + r\theta^2) = r(1 + r\theta^2)$ we find $\mathcal{I}(\theta) = -\mathbb{E}_\theta[h(\theta, Y)] = (r - r^2\theta^2)(1 + r\theta^2)^{-2} - r(1 + r\theta^2)(1 - 3r\theta^2)(1 + r\theta^2)^{-3} = 2r^2\theta^2(1 + r\theta^2)^{-2}$. Consequently, for $\theta > 0$,

$$T_n(\theta)^{1/2} = n^{-1/2} \sum_{i=1}^n \frac{s(\theta, Y_i)}{\sqrt{\mathcal{I}(\theta)}} = (2n)^{-1/2} \sum_{i=1}^n \left\{ -1 + \frac{(Y_i^\mathsf{T} 1_r)^2}{r(1 + r\theta^2)} \right\}.$$