Biometrika (0000), **000**, 0, *p*. 1
Printed in Great Britain Advance Access publication on 00 00 0000

An anomaly arising in the analysis of processes with more than one source of variability

BY H. S. BATTEY
Department of Mathematics, Imperial College London, London SW7 2AZ, UK
h.battey@imperial.ac.uk

PETER MCCULLAGH
Department of Statistics, University of Chicago, 5747 S Ellis Ave, Chicago, IL 60637, USA
pmcc@galton.uchicago.edu

SUMMARY

It is frequently observed in practice that the Wald statistic gives a poor assessment of the statistical significance of a variance component. This paper provides detailed analytic insight into the phenomenon by way of two simple models, which point to an atypical geometry as the source of the aberration. The latter can in principle be checked numerically to cover situations of arbitrary complexity, such as those arising from elaborate forms of blocking in an experimental context, or models for longitudinal or clustered data. The salient point, echoing Dickey (2020), is that a suitable likelihood-ratio test should always be used for the assessment of variance components.

Some key words: components of variance, likelihood ratio, nuisance parameters, REML, Wald statistic.

1. INTRODUCTION

Faithful representation of a process generating data often entails specification of two or more sources of variability. In an experimental context, simple or elaborate forms of blocking induce nested or crossed structure within the set of plots. Similar grouping arises in observational studies where, for instance, data may originate from different hospitals, regions, or from several family groups, of no direct interest but likely to generate structured correlation in the outcome.

As in certain other settings, inference based on the likelihood function for the full generative model is typically miscalibrated, sometimes seriously so, and should ideally be based on a suitable marginal or conditional likelihood. In the present context, appropriate preliminary reduction leads to residual maximum likelihood (REML), developed by Patterson and Thompson (1971), and closely connected to marginal likelihood (Bartlett, 1937).

Even when the REML likelihood is used, the Wald test routinely reported in software implementations is frequently found to be ineffectual for detecting components of variance when they are unambiguously present. A typical example occurs in chapter 4 of McCullagh (2023), where the likelihood-ratio statistic is more than eight times as large as the squared Wald ratio. The phenomenon has also been documented by Dickey (2020), who presented examples in which the REML Wald statistic is bounded. In at least one of the examples he considered, the natural estimate of standard error of the REML estimator is proportionate to the REML estimate itself, so that the role of the data in the Wald construction is negligible. The purpose of the present paper is to expose the matter in a form amenable to detailed analytic calculation, thereby revealing depen-

dence on key aspects and indicating other settings in which the same phenomenon is inevitable. Dickey's (2020) insights are reproduced and elucidated further. The source of the anomalous behaviour is not failure of distributional approximations obtained under hypothetical regimes of sometimes questionable adequacy, but rather atypical geometry of the (REML) log-likelihood function, which induces a bounded Wald statistic even under a notional limiting operation in which the likelihood-ratio statistic for testing the same hypothesis is arbitrarily large. We show that a version of the score statistic is subject to the same aberration.

The implications for applied work are consequential, as undetected sources of variability typically result in estimated standard errors for regression coefficient estimators that are deceptively small, and therefore confidence intervals that are misleadingly narrow. The opposite situation can occasionally arise as well. For instance, in a randomized blocks design, omission of the block factor as a variance component has the effect of increasing the estimated variance of treatment effect estimators.

2. VARIANCE COMPONENTS MODELS

In a typical variance-components model, the distribution of the response vector is specified by a mean vector $\mu = X\beta$ in the linear subspace $\mathcal{X}$ spanned by the columns of the model matrix X, and a covariance matrix Σ in the convex cone

$$\mathcal{V} = \Big\{ \Sigma = \sum_{u=0}^{s} \theta_u V_u : \theta_u \geq 0 \Big\}$$

spanned by given matrices $V_0, \ldots, V_s$, which are positive-definite or semi-definite. Usually $V_0 = I_n$ is the identity; the remaining matrices may be block factors or structured matrices associated with spatial or temporal dependence.

The residual likelihood is the likelihood function based on the residual $U^{\mathrm{T}}Y$, where $\ker(U^{\mathrm{T}}) = \mathcal{X}$. Provided that the matrices $U^{\mathrm{T}}V_uU$ are linearly independent, the variance components θ_u may be estimated by maximizing the residual likelihood. In subsequent discussion, reference to the log-likelihood function and its maximizer means the REML version unless otherwise specified.

There are compelling general arguments, notably invariance and permissibility of asymmetric confidence regions, for basing an assessment of the hypothesis $H_0 : \theta_s = 0$, say, on the likelihood-ratio statistic

$$\Lambda = 2\{\ell(\hat\theta) - \ell(\hat\theta^{(0)})\},$$

where $\hat\theta$ is the maximum likelihood estimator and $\hat\theta^{(0)}$ is the constrained estimator under H_0. Non-negativity of the variance coefficients means that the subset of $\mathcal{V}$ defined by the constraint $\theta_s = 0$ is a boundary sub-cone, with the implication that the likelihood achieves its maximum on the boundary with positive probability, usually one half for sufficiently large sample size (Chernoff, 1954). When an exact F statistic exists, the boundary event occurs whenever $F \leq 1$. On this event, $\hat\theta \in H_0$ coincides with $\hat\theta^{(0)}$ and the log-likelihood ratio is exactly zero. Thus with asymptotic probability one half $\Lambda = 0$; otherwise its distribution under H_0 is χ^2_1 under suitable limiting conditions. The realized value of Λ is to be calibrated against this distribution.

The nominal asymptotic variance of $\hat\theta_s$ is $i^{ss}(\theta)$, the (s,s) component of the inverse Fisher information matrix. With asymptotic probability one half, the squared Wald ratio $W^2 = \hat\theta_s^2 / i^{ss}(\hat\theta_s)$ is equal to zero under H_0 and otherwise it is asymptotically equivalent to Λ by a standard asymptotic argument. However, it is frequently observed in practice that W^2 gives a poor assessment

of the statistical significance of the sth variance component, in the sense that it fails to reject H_0 even when Λ does so unambiguously at the same significance level. If the likelihood is maximized on the boundary, both W^2 and Λ are zero and there is no disagreement. Disagreement can only occur when both are positive, and it is that phenomenon that we address here.

Among the substantial literature on testing variance components is an early contribution by Wald (1947), notable in that it recommends an F test but fails to warn against use of the eponymous statistic in this context. Subsequent contributions have developed non-standard asymptotic theory for the likelihood-ratio statistic and modifications thereof, relaxing an earlier assumption that the true parameter value belongs to the interior of the parameter space (e.g., Self and Liang, 1987; Geyer, 1994; Vu and Zhou, 1997). While the likelihood-ratio test and its variants have been the focus of theoretical development, common software implementations report Wald statistics as standard, without warning. Dickey (2020) appears to have been the first to emphasize the point at issue. The popularity of Wald-based inference perhaps stems from the convenience of its construction, requiring a single maximum likelihood fit in contrast with two for Λ, which facilitates the presentation of confidence statements.

The analysis of sections 3 and 6 is comparable to that of Dickey (2020), whose derivations cover situations in which the likelihood-ratio test coincides with an exact F test. The two papers illustrate the aberration under study in different ways and section 3.4 provides a comparison and synthesis. Sections 4 and 5 cover situations in which a fruitful formulation in terms of F may be infeasible but for which the shared anomalous geometry can be checked by direct study of the log-likelihood function. Together, the present paper and that of Dickey (2020) provide a thorough explanation of a phenomenon of broad relevance and scientific consequence.

3. Analysis for a single block factor

3.1. Introduction

We consider in this section a simple Gaussian model with two variance components estimated from the sufficient statistic, which consists of the within-blocks mean square MS_0 on f_0 degrees of freedom and the between-blocks mean square MS_1 on f_1 degrees of freedom. Specifically, the outcome is $Y_{ji} = \mu + \eta_j + \varepsilon_{ji}$, $(j = 1, \ldots, k,\ i = 1, \ldots, b)$ where, for an arbitrary block index j, $(Y_{j1}, \ldots, Y_{jb})^{\mathrm{T}}$ has a Gaussian distribution of mean $\mu 1_b$ and covariance matrix $\Sigma = \sigma_0^2 I_b + \sigma_\eta^2 1_b 1_b^{\mathrm{T}}$. Define, in a standard notation for averaging over suffixes, $\overline{Y}_{j\bullet} = \sum_i Y_{ji}/b$ and similarly for the double average. The between-blocks sum of squares

$$\sum_{j,i} (\overline{Y}_{j\bullet} - \overline{Y}_{\bullet\bullet})^2 = b \sum_j (\overline{Y}_{j\bullet} - \overline{Y}_{\bullet\bullet})^2 = f_1 \mathrm{MS}_1$$

is distributed as $\sigma_1^2 \chi^2_{f_1}$ where $f_1 = k - 1$ and, in the usual variance-components parameterization with $\theta \geq 0$,

$$\sigma_1^2 = \mathbb{E}(\mathrm{MS}_1) = b\sigma_\eta^2 + \sigma_0^2 = \sigma_0^2(1 + b\theta).$$

The within-blocks sum of squares $\sum_{j,i} (Y_{ji} - \overline{Y}_{j\bullet})^2 = f_0 \mathrm{MS}_0$ is $\sigma_0^2 \chi^2_{f_0}$ distributed independently of the between-blocks sum of squares, where $f_0 = k(b-1)$. The constraint on θ means that the null hypothesis $H_0 : \theta = 0$ of equality of variances is on the boundary.

In the balanced one-way analysis of variance structure of the above formulation, the REML log likelihood is the marginal log likelihood $\ell(\theta, \sigma_0^2)$ based on the joint density function of $(\mathrm{MS}_0, \mathrm{MS}_1)$. While numerous generalizations may be considered, the simple version presented

here isolates the point at issue in the most incisive form, free of secondary effects that complicate the analysis and interpretation.

3.2. *Comparison of Wald and likelihood ratio statistics*

Maximization of $\ell(\theta, \sigma_0^2)$ without the constraint $\hat\theta \geq 0$ produces estimators $\hat\theta = (F-1)/b$ and $\hat\sigma_0^2 = \mathrm{MS}_0$ where $F = \mathrm{MS}_1/\mathrm{MS}_0$ is Fisher's F ratio, whose distribution depends only on the variance ratio θ. The maximum likelihood estimator $\hat\theta$ has nominal asymptotic variance given by the relevant diagonal component of the inverse Fisher information matrix, namely

$$i^{\theta\theta}(\theta, \sigma_0^2) = i^{\theta\theta}(\theta) = \frac{2f(1+b\theta)^2}{f_0 f_1 b^2}, \tag{1}$$

where $f = f_1 + f_0$. The squared Wald statistic for testing $\theta = 0$ is therefore

$$W^2 = \hat\theta^2 \left\{ \frac{f_0 f_1 b^2}{2f(1+b\hat\theta)^2} \right\} = \Big(\frac{F-1}{F}\Big)^2 \frac{f_0 f_1}{2f},$$

to be compared with the likelihood ratio statistic

$$\begin{aligned} \Lambda &= f \log \mathrm{MS} - f_1 \log \mathrm{MS}_1 - f_0 \log \mathrm{MS}_0 \\ &= f \log(1 + f_1 b\hat\theta/f) - f_1 \log(1 + b\hat\theta), \end{aligned}$$

where the pooled mean square $\mathrm{MS} = (f_1 \mathrm{MS}_1 + f_0 \mathrm{MS}_0)/f$ is the maximum likelihood estimator of σ_0^2 under the constraint $\theta = 0$.

Although the statistics W^2 and Λ are known functions of $b\hat\theta = (F-1)$, simple approximations provide insight into the nature of the aberration described in section 1. Taylor expansion of W^2 and Λ around $\hat\theta = 0$ gives

$$\begin{aligned} W^2 &= \frac{b^2\hat\theta^2 f_0 f_1}{2f}(1 - 2b\hat\theta) + O(\hat\theta^4) \\ \Lambda &= \frac{b^2\hat\theta^2 f_0 f_1}{2f}\left\{1 - \frac{2b\hat\theta(f_0 + 2f_1)}{3f}\right\} + O(\hat\theta^4) \\ \Lambda/W^2 &= 1 + \frac{2b\hat\theta(2f_0 + f_1)}{3f} + \frac{4b^2\hat\theta^2(2f_0 + f_1)}{3f} + O(\hat\theta^3), \end{aligned} \tag{2}$$

or in terms of $\Lambda \geq 0$

$$\Lambda/W^2 = 1 + \frac{2\sqrt{2}(2f_0 + f_1)\Lambda^{1/2}}{3(f f_0 f_1)^{1/2}} + \frac{4(2f_0 + f_1)(3f + f_1)\Lambda}{3 f f_0 f_1} + O(\Lambda^{3/2}),$$

showing that they agree up to second order in $\hat\theta$ and to first order in Λ, but not beyond. Typically $f_0 \gg f_1$ is large, in which case the ratio is approximately

$$\Lambda/W^2 \simeq 1 + \frac{4b\hat\theta}{3} = 1 + \frac{4(F-1)}{3} \simeq 1 + \frac{4\sqrt{2\Lambda/f_1}}{3}, \quad \Lambda \geq 0 \tag{3}$$

for small $\hat\theta$ or Λ.

The squared Wald statistic W^2 is justified based on standard asymptotic theory for maximum likelihood estimators. First- and second-moment theory suggests three further Wald statistics, for which the same anomalous behaviour is demonstrated in S.2 of the supplementary material. For a similar discussion of two score statistics, see S.3.

In the simplest setting, at least, both W and $\Lambda^{1/2}\text{sign}(F-1)$ are monotone functions of the mean-square ratio F, so their distributions are known exactly as a function of the variance-ratio parameter θ. The transformations $\Lambda = g_1(F)$ and $W^2 = g_2(F)$ are strictly monotone increasing for $F > 1$, and decreasing for $F < 1$. The discrepancies described here arise from the standard practice of treating Λ and W^2 as if they were asymptotically identical statistics rather than equivalent statistics. The standard practice is justified only if $g_1 = g_2$, at least approximately for large samples, which is not the case in typical variance-components models.

One definition of Wald-detectability equates θ to $\Phi^{-1}(1-\alpha)$ times the estimated standard error of $\hat\theta$. It can be shown (supplementary material S.2) that there does not exist a positive Wald-detectable value in this sense unless the number of blocks is large. It is arguably more natural to compute standard errors under the null, in which case the values at the borderline of Wald-detectability are

$$\theta^*_{16} = \left(\frac{2f}{f_0 f_1 b^2}\right)^{1/2}, \quad \theta^*_2 = 2\theta^*_{16}$$

at the 16% and 2% levels. The corresponding thresholds in terms of F are

$$F^*_{16} = 1 + \left(\frac{2f}{f_0 f_1}\right)^{1/2}, \quad F^*_2 = 2F^*_{16}.$$

With $f_0 = 102$, $f_1 = 5$ and $b = 18$, the ratio Λ/W^2 is approximately 1.864 at θ^*_{16} and 2.727 at θ^*_2. With the same numbers, $F^*_{16} = 1.648$ and $F^*_2 = 3.296$, to be compared with the 16% and 2% critical values of the F distribution on (f_1, f_0) degrees of freedom, which are 1.625 and 2.818 respectively. Thus, even when standard errors are computed under the null hypothesis, the observed value of the F statistic has to be 17% larger than the 2% critical value of the F distribution in order for the Wald test to reject at the same level. Equivalently, rejection at the 2% level using a Wald statistic with standard errors computed under the null requires an observed value of $\hat\theta$ that is double that necessary for rejection at the same level using an F test. These latter conclusions do not involve any approximations, and are similar and complementary to those of Dickey (2020).

Geometric insight is obtained by noting that the Wald statistic W^2 implicitly defines a quadratic approximation $q(\theta)$ to the profile REML log likelihood function, $\ell(\theta; \hat\sigma^2_\theta)$ in a neighbourhood of $\hat\theta = (F-1)/b$. Specifically, for fixed θ, the maximum likelihood estimator of σ^2_0 is $\hat\sigma^2_\theta = w_0\text{MS}_0 + w_1\text{MS}_1/(1+b\theta)$ where $w_r = f_r/f$ and

$$2\ell(\theta; \hat\sigma^2_\theta) = -f[1 + \log\{f_0\text{MS}_0(1+b\theta) + f_1\text{MS}_1\}] + f_0\log(1+b\theta). \tag{4}$$

Write $\ell(\theta; \hat\sigma^2_\theta) = \hat\ell(\theta) + f/2$. The quadratic approximation to $\hat\ell(\theta)$ implicit in the Wald test is

$$q(\theta) = -\left(\frac{f_0 f_1 b^2}{2fF^2}\right)(\theta - \hat\theta)^2 + \hat\ell(\hat\theta),$$

where $f_0 f_1 b^2/(2fF^2)$ is $1/i^{\theta\theta}(\hat\theta)$ and

$$2\hat\ell(\hat\theta) = -f\log f - f_1\log\text{MS}_1 - f_0\log\text{MS}_0.$$

Since the two functions have the same value at $\hat\theta$, the discrepancy between the Wald and likelihood ratio statistics for testing $\theta = 0$ is the difference in y-intercepts:

$$q(0) - \hat\ell(0) = \frac{\Lambda}{2} - \left(\frac{f_0 f_1}{2f}\right)\left(\frac{F-1}{F}\right)^2, \tag{5}$$

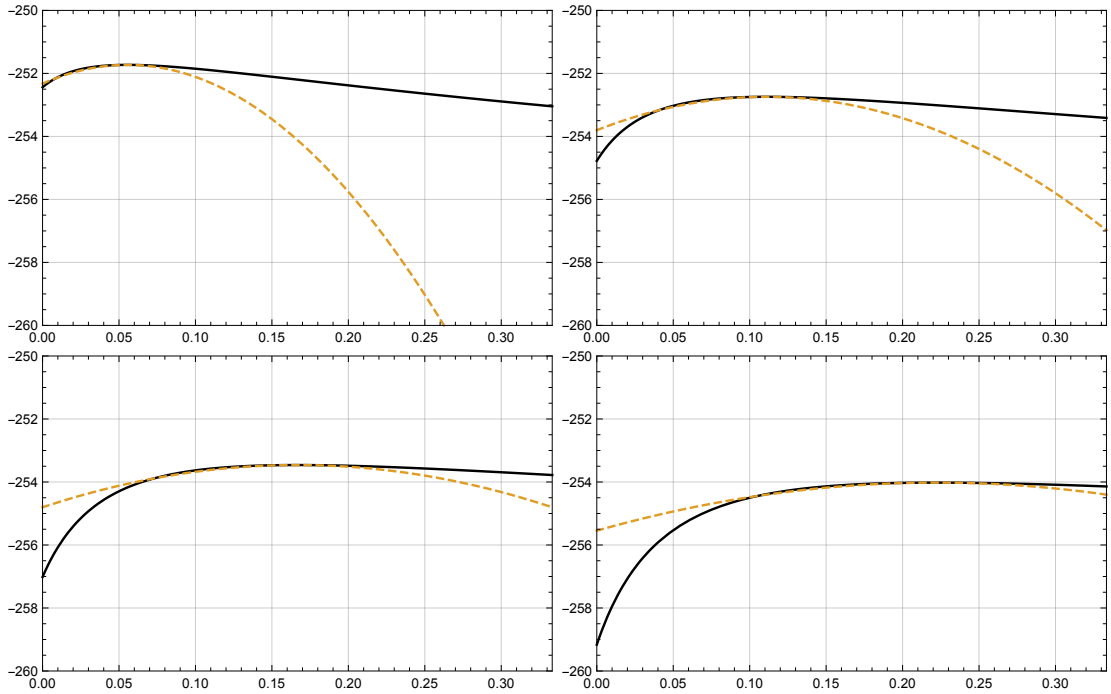

Fig. 1. Graphs of the profile REML log likelihood function $\hat{\ell}(\theta)$ (black) and its quadratic approximation $q(\theta)$ (orange dashed) for $f_1 = 5$, $f_0 = 102$, $b = 18$, $\mathrm{MS}_0 = 1$ and (from top left to bottom right) $F \in \{2, 3, 4, 5\}$.

where the expression for Λ in terms of F is

$$\Lambda = f \log(1 + f_1(F-1)/f) - f_1 \log F.$$

For F and f_1 fixed,

$$q(0) - \hat{\ell}(0) = \frac{f_1}{2}\Big\{(F-1)\Big(\frac{F^2 - F + 1}{F^2}\Big) - \log F\Big\} + O(f_0^{-1}),$$

showing that the discrepancy is roughly linear in F for large f_0, while for fixed f_1 and f_0, (5) converges to zero as F approaches unity and is unbounded for F arbitrarily large. Fig. 1 graphs $q(\theta)$ and $\hat{\ell}(\theta)$ for different values of F.

3.3. *Non-constant Fisher information and anomalous geometry*

From equation (4), $\hat{\ell}(\theta)$ has second derivative

$$\gamma(\theta) = \frac{b^2 f_0}{2}\left\{\frac{f_0 f \mathrm{MS}_0^2}{(f_1 \mathrm{MS}_1 + f_0 \mathrm{MS}_0(1 + b\theta))^2} - \frac{1}{(1+b\theta)^2}\right\},$$

whose value at $\hat{\theta}$ is

$$\gamma(\hat{\theta}) = -\frac{b^2 f_0 f_1}{2F^2} = -1/i^{\theta\theta}(\hat{\theta}), \tag{6}$$

showing that the curvature at the maximum-likelihood point is close to zero for large F, as depicted in Fig. 1. In other words, $\hat{\ell}(\theta)$ is arbitrarily well approximated by a horizontal asymptote

at $\hat\ell(\hat\theta)$ in a neighbourhood of $\hat\theta$ for arbitrarily large F. Equation (6) also shows that the discrepancy $q(0)-\hat\ell(0)$ is attributable to the higher-order derivatives, since there is no error incurred by using $-1/i^{\theta\theta}(\hat\theta)$ in place of $\gamma(\hat\theta)$ in the Taylor series approximation to $\hat\ell(\theta)$. The effect of higher-order derivatives is encapsulated to a large extent in the considerable non-constancy of $i^{\theta\theta}(\theta)$ as a function of θ over the range of interest.

From (1), the ratio of the nominal asymptotic variances of $\hat\theta$ at arbitrary θ and at $\theta=0$ is $i^{\theta\theta}(\theta)/i^{\theta\theta}(0)=(1+b\theta)^2$, and the range of primary interest is

$$0\le\theta\le\sqrt{2f/b^2f_0f_1},$$

the upper value being the nominal asymptotic standard deviation of $\hat\theta$ under the null hypothesis, having conditioned on the event that $\hat\theta$ is not on the boundary. Over this range, the asymptotic variance varies by a factor

$$1\le(1+b\theta)^2\le(1+\sqrt{2f/f_0f_1})^2,$$

which is large for typical values of f_1. For example, $f_0=102$, $f_1=5$, gives a range of approximately $1\le(1+b\theta)^2\le 2.72$.

3.4. A synthesis with Dickey (2020)

The motivation for the present paper came from practical examples in chapter 4 of McCullagh (2023), where the Wald statistic was ineffectual at detecting variance components, the absence of which was strongly refuted by a likelihood ratio test. Dickey (2020) exposed the same phenomenon. His exposition, aimed at practitioners, covers widely used models in the analysis of designed experiments for which exact F and Wald tests are available. The present paper is framed in terms of the log-likelihood ratio statistic Λ, which is more generally available and points to geometric insights not recoverable from the moment-based Wald constructions. Three instances of the latter are discussed in the supplementary material, among which equation (S.1) coincides with equation (2) of Dickey (2020). For cases where F is available, the present paper and that of Dickey provide equivalent explanations from two points of view.

4. TWO-SAMPLE PROBLEM IN GENERAL SCALE FAMILIES

Let Y be a random variable from a scale family with density function $\tau^{-1}g(y/\tau)$, $y,\tau>0$, where g is a known, continuous, density function on the positive real line. Let

$$I_g=\int_0^\infty\frac{\{xg'(x)+g(x)\}^2}{g(x)}\,dx.$$

Then the Fisher information for τ in an independent and identically distributed sample $Y_1,\ldots,Y_n$ is nI_g/τ^2 and that for τ_0,τ_1 in a two-sample problem of sizes n_0 and n_1 is

$$I_g\,\mathrm{diag}\Big(\frac{n_0}{\tau_0^2},\frac{n_1}{\tau_1^2}\Big). \tag{7}$$

The squared Wald statistic for testing equality of scale parameters is therefore

$$W^2=\frac{n_0n_1I_g(\hat\tau_1-\hat\tau_0)^2}{n_0\hat\tau_1^2+n_1\hat\tau_0^2}=\frac{n_0n_1I_g(\hat\theta-1)^2}{n_0\hat\theta^2+n_1},$$

where $\hat\theta$ is the estimated ratio $\hat\tau_1/\hat\tau_0$. The Wald statistic is bounded in both limits for $\hat\theta$ arbitrarily large or small. Specifically

$$\lim_{\hat\theta\to 0} W^2 = n_0 I_g, \quad \lim_{\hat\theta\to\infty} W^2 = n_1 I_g.$$

The curvature of the profile log-likelihood function at $\hat\theta$ is approximated by the expected Fisher information in the profile log likelihood. The Fisher information transforms as

$$i^{(\psi)}_{ab}(\psi) = \frac{\partial\phi^r}{\partial\psi^a}\frac{\partial\phi^s}{\partial\psi^b} i^{(\phi)}_{rs}(\phi), \tag{8}$$

where ψ and ϕ are two parameterizations and we have used the convention that symbols appearing both as subscripts and superscripts in the same product are summed. The information about θ, having adjusted for estimation of τ_0 is therefore, using (7) and (8),

$$i_{\theta\theta.\tau_0} = i_{\theta\theta} - i^2_{\theta\tau_0}/i_{\tau_0\tau_0} = \frac{n_0 n_1}{\theta^2(n_0+n_1)} I_g,$$

showing that the two-sample problem in general scale families has the same anomalous geometry documented in section 3 for large θ, leading to the discrepancy between the likelihood-ratio and the Wald statistics.

5. GAUSSIAN VARIANCE COMPONENTS MODEL

Consider a variance-components model in which $Y \in \mathbb{R}^n$ is normal with mean $\mu \in \mathcal{X}$ of dimension $p < n$, and covariance matrix $\Sigma = \sigma^2(I_n + V(\theta))$ where $V(\theta)$ is a known matrix function of a vector parameter $\theta = (\theta_1, \ldots, \theta_s)^{\mathrm{T}}$ with $V(0) = 0$. This encompasses the linear model $V(\theta) = \sum_u \theta_u V_u$ from section 2. The subspace $\mathcal{X} \subset \mathbb{R}^n$ is a group under addition, which implies that for any matrix U^{T} with kernel $\mathcal{X}$, the normalized residual statistic

$$q(y, U) = U^{\mathrm{T}}y/\|U^{\mathrm{T}}y\| = U^{\mathrm{T}}y/(y^{\mathrm{T}}UU^{\mathrm{T}}y)^{1/2}. \tag{9}$$

is maximal invariant under the affine group with action $g(a, x) : y \mapsto ay + x$, with $a > 0$, and $x \in \mathcal{X}$. For distributional calculations, it is convenient to take the columns of U to be an orthonormal basis in $\mathcal{X}^\perp$, the orthogonal complement of $\mathcal{X}$ with respect to the standard inner product. In that case $UU^{\mathrm{T}} = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, $U^{\mathrm{T}}U = I_{n-p}$, and the density function of $Q = q(Y, U)$ is (supplementary material S.4)

$$\frac{\Gamma\big(\frac{n-p}{2}\big)|A_\theta|^{-1/2}(q^{\mathrm{T}}A_\theta^{-1}q)^{-(n-p)/2}}{2\pi^{(n-p)/2}} dq_1 \cdots dq_{n-p}, \tag{10}$$

where $\sigma^2 A_\theta = U^{\mathrm{T}}\Sigma U$. At $\theta = 0$, Q is uniformly distributed on the $(n-p)$-dimensional unit sphere in $\mathbb{R}^n$. The above exposition hybridizes Kariya (1980) and King (1980).

By construction, the distribution (10) does not depend on β or σ^2, and inference for θ is conveniently based on the marginal log likelihood function

$$\breve\ell(\theta) = -\tfrac{1}{2}\log|A_\theta| - \tfrac{(n-p)}{2}\log(q^{\mathrm{T}}A_\theta^{-1}q), \tag{11}$$

closely related to the REML log likelihood function which uses the marginal distribution of $U^{\mathrm{T}}Y$ rather than Q. Since the transformation to Q eliminates the nuisance parameter σ^2 as well as β, analysis based on (11) is more amenable to analytic calculation.

The density function in (10) is relative to a particular orthonormal basis in $\mathcal{X}^\perp$, and the same basis is embedded in the likelihood function (11) in the matrix A_θ. The conclusion in (11) is

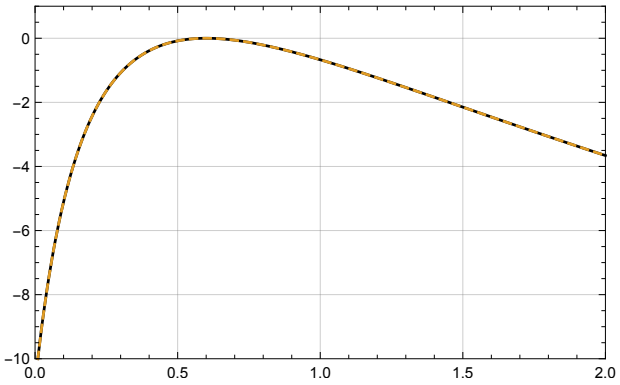

Fig. 2. Graph of $\hat{\ell}(\theta) - \hat{\ell}(\hat{\theta})$ from section 3 (black) and $\breve{\ell}(\theta) - \breve{\ell}(\breve{\theta})$ from equation (12) (orange dashed) for $f_1 = 22$, $f_0 = 92$, $b = 5$, $\mathrm{MS}_0 = 1$ and $F = 4$.

equivalent to equation (3.3) McCullagh (2009), which evades the problem of selecting a basis by allowing singular matrices.

The single block factor setting of section 3 is a special case with $n = (f_1 + 1)b$, $\Sigma = \sigma^2(I_n + \theta V)$, $V = I_{f_1+1} \otimes 1_b 1_b^{\mathrm{T}}$ and $\mathcal{X} = 1$, so that $UU^{\mathrm{T}} = I_n - 1_n 1_n^{\mathrm{T}}/n$. Suppose for an exact analytic calculation that $f_1 + 1$ and b are both powers of two. Supplementary material S.5 shows that $\log|A_\theta| = f_1 \log(1 + b\theta)$ and

$$q^{\mathrm{T}} A_\theta^{-1} q = 1 + \textstyle\sum_{j=1}^{f_1} q_{jb}^2 \{(1 + b\theta)^{-1} - 1\}$$

so that (11) becomes

$$\begin{aligned} \breve{\ell}(\theta) &= -\frac{f_1}{2}\log(1 + b\theta) - \frac{f}{2}\log\left\{1 - \frac{b\theta \sum_{j=1}^{f_1} q_{jb}^2}{(1 + b\theta)}\right\} \\ &= -\frac{f_1}{2}\log(1 + b\theta) - \frac{f}{2}\log\left\{1 - \frac{b\theta f_1 F}{(1 + b\theta)(f_0 + f_1 F)}\right\}. \end{aligned} \tag{12}$$

The solution of the likelihood equation by differentiation of (12) is $(1 + b\breve{\theta}) = F$ and direct calculation shows that $\breve{\ell}(\theta) - \breve{\ell}(\breve{\theta}) = \hat{\ell}(\theta) - \hat{\ell}(\hat{\theta})$ from (4). This is demonstrated empirically in Fig. 2 with the values of f_0, f_1 and b from the numerical example of section 6.1 so as to also illustrate the anomalous geometry. When an analysis of variance is feasible, it produces identical estimates to those based on maximization of (11), the only difference arising from the choice of basis for the orthogonal subspaces.

More generally, in a model with a scalar parameter θ generating a variance component in the form $\Sigma = \sigma^2(I_n + V(\theta))$, the likelihood equation for θ is

$$\frac{\mathrm{tr}(A_\theta^{-1}\dot{A}_\theta)}{(n - p)} = \frac{q^{\mathrm{T}}(A_\theta^{-1}\dot{A}_\theta A_\theta^{-1})q}{q^{\mathrm{T}} A_\theta^{-1} q}$$

at $\theta = \breve{\theta}$, where $\dot{A}_\theta = \nabla_\theta A_\theta$ and by the Woodbury identity

$$A_\theta^{-1} = I_{n-p} - U^{\mathrm{T}}(V(\theta)^{-1} + UU^{\mathrm{T}})^{-1}U.$$

The second derivative of $\breve{\ell}(\theta)$ at an arbitrary point is

$$\breve{\gamma}(\theta) = \nabla^2_{\theta\theta}\breve{\ell}(\theta) = -\tfrac{1}{2}\mathrm{tr}\big(A_\theta^{-1}\ddot{A}_\theta - A_\theta^{-1}\dot{A}_\theta A_\theta^{-1}\dot{A}_\theta\big) - \frac{(n-p)}{2}\left\{\frac{q^{\mathrm{T}}A_\theta^{-1}(2\dot{A}_\theta A_\theta^{-1}\dot{A}_\theta - \ddot{A}_\theta)A_\theta^{-1}q}{q^{\mathrm{T}}A_\theta^{-1}q} + \frac{q^{\mathrm{T}}(A_\theta^{-1}\dot{A}_\theta A_\theta^{-1})q}{(q^{\mathrm{T}}A_\theta^{-1}q)^2}\right\},$$

where $\ddot{A}_\theta = \nabla^2_{\theta\theta}A_\theta$. A general analytic approximation to the curvature $\breve{\gamma}(\breve{\theta})$ has not been ascertained. However, if $V(\theta)$ is of the form θV with V a known matrix, then $\dot{A}_\theta = U^{\mathrm{T}}VU$, $\ddot{A}_\theta = 0$ and the previous display simplifies. In particular

$$\lim_{\theta\to\infty}\mathrm{tr}\big(A_\theta^{-1}\ddot{A}_\theta - A_\theta^{-1}\dot{A}_\theta A_\theta^{-1}\dot{A}_\theta\big) = 0,$$

$$\lim_{\theta\to\infty} A_\theta^{-1}\dot{A}_\theta A_\theta^{-1} = 0,$$

and

$$\lim_{\theta\to\infty} A_\theta^{-1}\big(2\dot{A}_\theta A_\theta^{-1}\dot{A}_\theta - \ddot{A}_\theta\big)A_\theta^{-1} = 0,$$

so that $\lim_{\theta\to\infty}\nabla^2_{\theta\theta}\breve{\ell}(\theta) = 0$. By consistency of the maximum likelihood estimator, the marginal log likelihood function has zero curvature at the maximising point when the true value of θ is arbitrarily large.

6. NUMERICAL ILLUSTRATIONS

6.1. *Covariance determined by split-plot nested blocking*

The split-plot covariance $\sigma_0^2(I_n + \theta V)$ is a linear combination of the identity and a binary matrix such that $V_{ij} = 1$ if observational units i, j belong to the same whole-plot or block. Chapter 1 of McCullagh (2023) discusses an example of this type with $l = 24$ rats constituting the blocks, and $s = 5$ sites on each rat constituting the observational units. A two-level treatment was assigned at random to the rats, so the model formula `site+treat` for the expected value determines a subspace of dimension 6. This is not a split-plot design in the traditional sense because the split-plot effect is associated with a classification factor (sites ranging from anterior to caudal), not with a treatment factor.

The real experiment is a little more complicated because some components are missing. With this exception, the simulated data mirror that example, and illustrate the effect of increasing θ on the discrepancy between Λ and W^2. Since $\hat{\theta} \simeq 0.4$ for the actual experiment, the range $0 \le \theta \le 1$ was used for simulation. For each of 1000 Monte Carlo replications, outcomes were generated according to the model $Y \sim N(\mu, \Sigma)$, the particular point $\mu \in \mathcal{X}$ being immaterial for the REML likelihood. As it happens, the analysis of section 3 applies here with two modified mean squares, one for treatment and one for residuals eliminating additive row and column effects, namely

$$f_0\mathrm{MS}_0 = \|(I_n - P_{\text{rat}} - P_{\text{site}} + P_0)Y\|^2 \sim \sigma_0^2\chi^2_{f_0},$$
$$f_1\mathrm{MS}_1 = \|(P_{\text{rat}} - P_{\text{trt}})Y\|^2 \sim \sigma_1^2\chi^2_{f_1},$$

where $f_1 = l - 2$, $f_0 = (l-1)(s-1)$ and, for example, $P_{\text{trt}}Y$ is the projection of Y on the subspace spanned by the treatment basis. For the null hypothesis $\theta = 0$, Fig. 3 compares the simulated values of Λ/W^2 with the linear approximation $1 + 4s\hat{\theta}/3$ from (3).

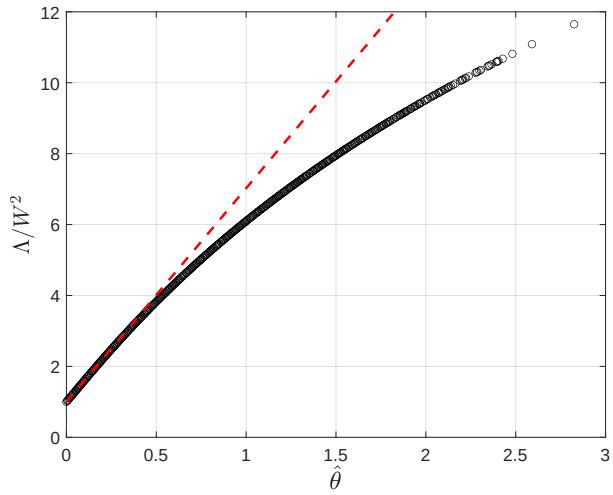

Fig. 3. Simulated Λ/W^2 plotted against $\hat{\theta}$ for the nested-blocks arrangement with $l = 24$ large blocks and $s = 5$ nested blocks.

Maximization of the marginal log-likelihood function $\breve{\ell}(\theta)$ from equation (11) is equivalent. Both normed log-likelihood functions are graphed in Fig 2 for $\hat{\theta} = \breve{\theta} = 0.6$, from which the anomalous geometry is apparent.

6.2. *Covariance determined by Latin square blocking*

The Latin square covariance $\Sigma = \sigma_0^2(I_n + \theta_1 V_1 + \theta_2 V_2)$ is a linear combination of the identity and two binary matrices of the form $(V_1)_{ij} = 1$ if observational units i and j share a column and $(V_2)_{ij} = 1$ if they share a row. For the simulation, the variance component parameters were taken as $\sigma_0^2 = 1$, $\theta_2 = 0.2$, while θ_1 was varied in the range $0 \le \theta_1 \le 1$. The relevant mean squares on $f_0 = (b-1)(b-2)$ and $f_1 = (b-1)$ degrees of freedom are

$$f_0 \mathrm{MS}_0 = \|(I_n - P_{\mathrm{col}} - P_{\mathrm{row}} - P_{\mathrm{trt}} + 2P_0)Y\|^2 \sim \sigma_0^2 \chi^2_{f_0},$$
$$f_1 \mathrm{MS}_1 = \|(P_{\mathrm{col}} - P_0)Y\|^2 \sim \sigma_1^2 \chi^2_{f_1},$$

where $\sigma_1^2 = \sigma_0^2(1 + b\theta_1)$. The analogous mean square MS_2 for rows on $f_2 = f_1$ degrees of freedom is used in variance estimation. Specifically, the information about θ_1, having adjusted for estimation of σ_0^2 and θ_2 is

$$\frac{b^2 f_1}{2(1+b\theta_1)^2}\left\{1 - \frac{f_1^2(1+b\theta_2)^2}{f_2 f(1+b\theta_1)^2}\right\}. \tag{13}$$

In (13), θ_j is estimated as the positive part of $(F_j - 1)/b$ with $F_j = \mathrm{MS}_j/\mathrm{MS}_0$. For $f = f_1 + f_0 \gg f_1$, which amounts to b being large, the adjustment is negligible provided that θ_2 is not too large relative to θ_1, and (13) is comparable to (1). The resulting Wald statistic for testing the hypothesis $\theta_1 = 0$ is to be compared with Λ, whose form is identical to that of section 3.

Fig. 4 depicts qualitatively similar behaviour for the ratio Λ/W^2 as that of Fig. 3, with additional variability attributable to randomness in $\hat{\theta}_2$ manifesting through the estimate of (13). The version with θ_2 treated as known is depicted in the right panel of Fig. 4.

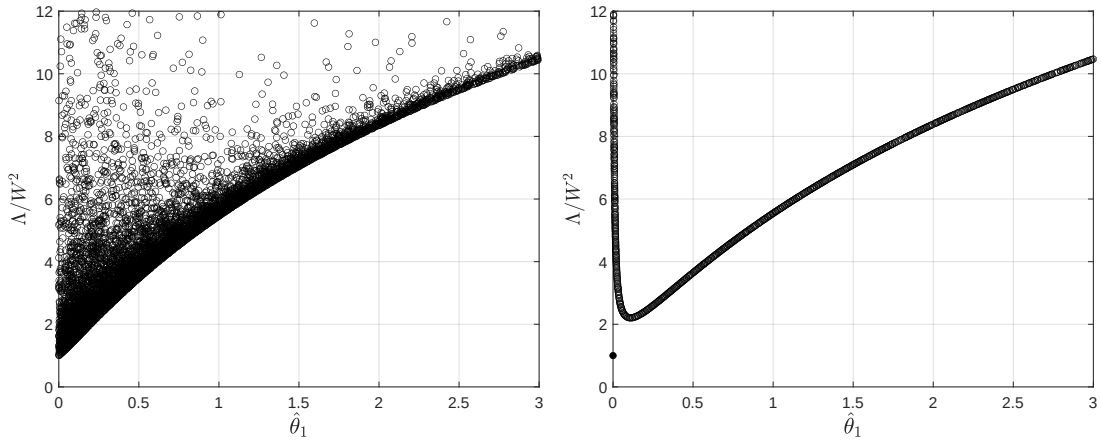

Fig. 4. Simulated Λ/W^2 plotted against $\hat{\theta}_1$ for the Latin square with $b = 6$ rows, columns and treatments. Left: θ_2 estimated in (13); right: θ_2 treated as known.

Acknowledgements

Section 3.4 was based on a detailed comparison supplied by one of two anonymous referees, to whom we are grateful.

Supplementary material

Supplementary material provides more explicit derivations for some of the key results, together with a discussion of three further Wald statistics and two score statistics.

REFERENCES

[1] BARTLETT, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. A*, 155, 268–282.
[2] CHERNOFF, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.*, 25, 573–578.
[3] DICKEY, D. A. (2020). A warning about Wald tests. *SAS Global Forum*, paper 5088 - 2020.
[4] GEYER, C. J. (1994). On the asymptotics of constrained M-estimation. *Ann. Statist.*, 22, 1993–2010.
[5] KARIYA, T. (1980). Locally robust tests for serial correlation in least squares regression. *Ann. Statist.*, 8, 1065–1070.
[6] KING, M. L. (1980). Robust tests for spherical symmetry and their application to least squares regression. *Ann. Statist.*, 8, 1265–1271.
[7] MCCULLAGH, P. (2009). Marginal likelihood for distance matrices. *Statist. Sinica*, 19, 631–649.
[8] MCCULLAGH, P. (2023). *Ten Projects in Applied Statistics*, Springer.
[9] PATTERSON, H. D. and THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545–554.
[10] SELF, S. G. and LIANG, K-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions *J. Amer. Statist. Assoc.*, 82, 605–610.
[11] VU, H. T. V. and ZHOU, S. (1997). Generalization of likelihood ratio tests under nonstandard conditions. *Ann. Statist.*, 22, 1993–2010.
[12] WALD, A. (1947). A note on regression analysis. *Ann. Math. Statist.*, 18, 586–589.

[*Received on 26 January 2023. Editorial decision on 00* 000 *0000*]

Biometrika (0000), **000**, 0, *p*. 1
Printed in Great Britain Advance Access publication on 00 00 0000

Supplementary material for 'An anomaly arising in the analysis of processes with more than one source of variability'

BY H. S. BATTEY
Department of Mathematics, Imperial College London, London SW7 2AZ, UK
h.battey@imperial.ac.uk

PETER MCCULLAGH
Department of Statistics, University of Chicago, 5747 S Ellis Ave, Chicago, IL 60637, USA
pmcc@galton.uchicago.edu

SUMMARY

Supplementary material provides more explicit derivations for some of the key results, together with a discussion of three further Wald statistics and two score statistics.

S.1. INVERSION OF THE TAYLOR EXPANSION OF Λ

From section 3.2, Taylor expansion of Λ shows that to a first approximation $\hat\theta^2 = 2f\Lambda/b^2 f_0 f_1$. Thus, write $\hat\theta_0 = b^{-1}(2f\Lambda/f_0f_1)^{1/2}$ and expand Λ around $\hat\theta = 0$ in the form

$$\Big[\frac{2f}{b^2 f_0 f_1}\Big\{f\log(1+f_1 b\hat\theta/f) - f_1\log(1+b\hat\theta)\Big\}\Big]^{1/2} = \hat\theta - \frac{b(f+f_1)\hat\theta^2}{3f} + O(\hat\theta^3).$$

Equating the right hand side to $\hat\theta_0$ and inverting for $\hat\theta$ gives

$$\begin{aligned}\hat\theta &= \hat\theta_0 + \frac{b(f+f_1)\hat\theta_0^2}{3f} + \frac{2b^2(f+f_1)^2\hat\theta_0^3}{3^2f^2} + O(\hat\theta_0^4)\\ &= \frac{\sqrt{2f\Lambda/f_0f_1}}{b} + \frac{2(f+f_1)\Lambda}{3bf_0f_1} + O(\Lambda^{3/2}),\end{aligned}$$

and by a similar inversion for $\hat\theta^2$,

$$\hat\theta^2 = \frac{2f\Lambda}{b^2f_0f_1} + O(\Lambda^{3/2}).$$

On using these expressions in the order $O(\hat\theta^3)$ expansion for Λ/W^2,

$$\Lambda/W^2 = 1 + \frac{2\sqrt{2}(2f_0+f_1)\Lambda^{1/2}}{3(ff_0f_1)^{1/2}} + \frac{4(2f_0+f_1)(3f+f_1)\Lambda}{3ff_0f_1} + O(\Lambda^{3/2}).$$

S.2. THREE FURTHER WALD STATISTICS

Depending on the chosen parameterization, the natural variance component estimate may be constructed from a difference of mean squares, a ratio of mean squares or the log-ratio of mean squares. First and second moment theory thus suggests three further Wald statistics appropriate

in some contexts and included here for comparison. The likelihood ratio statistic is invariant to the chosen parameterization.

The first version of the Wald statistic is based on the difference of mean squares $\mathrm{MS}_1 - \mathrm{MS}_0$, whose variance is

$$\mathrm{var}(\mathrm{MS}_1 - \mathrm{MS}_0) = \frac{2\sigma_0^4}{f_0} + \frac{2\sigma_1^4}{f_1}.$$

The estimated variance of the difference is

$$\frac{2\hat\sigma_0^4}{f_0} + \frac{2\hat\sigma_1^4}{f_1} = \frac{2\mathrm{MS}_0^2}{f_0} + \frac{2\mathrm{MS}_1^2}{f_1}.$$

In keeping with section 3, the variance of $\mathrm{MS}_1 - \mathrm{MS}_0$ is computed and estimated for general θ rather than $\theta = 0$. Thus, the squared Wald statistic simplifies to

$$W_1^2 = \frac{(\mathrm{MS}_1 - \mathrm{MS}_0)^2}{\frac{2\mathrm{MS}_0^2}{f_0} + \frac{2\mathrm{MS}_1^2}{f_1}} = \frac{(F-1)^2}{2}\frac{f_0 f_1}{f_1 + F^2 f_0}, \tag{S.1}$$

which is equation (2) of Dickey (2020). Evidently, W_1^2, like W^2 and the other versions to be presented, is scale invariant and has a distribution depending only on the variance ratio.

The second version of the Wald statistic is a linear function of the sample F-ratio $F = \mathrm{MS}_1/\mathrm{MS}_0$, whose distribution is $(1 + b\theta)F_{f_1,f_0}$, a scalar multiple of Fisher's F distribution on f_1, f_0 degrees of freedom. Provided that the within-blocks degrees of freedom satisfies $f_0 > 4$, the variance of F is

$$\mathrm{var}(F) = \frac{\sigma_1^4}{\sigma_0^4} \times \frac{2f_0^2(f-2)}{f_1(f_0-2)^2(f_0-4)},$$

and since $\hat\sigma_1^2/\hat\sigma_0^2 = F$, the estimated variance is

$$F^2 \times \frac{2f_0^2(f-2)}{f_1(f_0-2)^2(f_0-4)}.$$

The second version of the Wald statistic is therefore

$$W_2^2 = \Big(\frac{F-1}{F}\Big)^2 \frac{f_1(f_0-2)^2(f_0-4)}{2f_0^2(f-2)} \simeq \Big(\frac{F-1}{F}\Big)^2 \frac{f_0 f_1}{2f} = W^2. \tag{S.2}$$

The approximation is asymptotic for fixed f_1 and large f_0.

The third version of the Wald statistic is a linear function of the log F-ratio, whose variance is

$$\mathrm{var}\log F = \psi'(f_0/2) + \psi'(f_1/2),$$

where ψ is the derivative of the log gamma function, satisfying $\psi'(x) = 1/x + O(1/x^2)$ for large x. On the log scale, the variance is constant and independent of the parameter, so no data-dependent estimate is needed. Thus, the third version for comparison is

$$W_3^2 = \frac{\log^2 F}{\psi'(f_0/2) + \psi'(f_1/2)} \simeq \log^2 F \times \frac{f_0 f_1}{2f} \tag{S.3}$$

In this case, the approximation requires both degrees of freedom to be large.

The parameter constraint $\theta \geq 0$ or $\sigma_1 \geq \sigma_0$, can be imposed by replacing F by $\max(F, 1)$ in (S.1)-(S.3).

For $f_0 \gg f_1$, $W_1^2 \simeq W^2$, while for any f_0, f_1, $W^2 \leq W_1^2 \leq f_1/2$. Thus, W_1^2 is subject to the same deficiency as W^2. Equation (S.2) shows that W_2^2 is similarly problematic. The anomalous

behaviour persists for all θ and the variance component is not detectable at any threshold unless the number of blocks is large.

For small $\hat\theta$

$$\log^2(1+b\hat\theta)\times\frac{f_0f_1}{2f}=\frac{\hat\theta^2b^2f_0f_1}{2f}(1-b\hat\theta)+O(\hat\theta^4). \tag{S.4}$$

The Taylor expansion around $\hat\theta=0$ of the ratio of Λ to equation (S.4) is

$$1+\frac{b\hat\theta(f_0-f_1)}{3f}+\frac{b^2\hat\theta^2(f_0-f_1)}{3f}+O(\hat\theta^3)$$

and therefore, approximately for small $\Lambda\geq 0$,

$$\Lambda/W_3^2\simeq 1+\frac{\sqrt{2f\Lambda/f_0f_1}}{3}, \tag{S.5}$$

to be compared with (3). The approximation (S.5) requires $f_0\gg f_1$, both large.

S.3. Two score statistics

Two versions of the score statistic are

$$S_1^2=\nabla_\theta\hat\ell(0)^2i^{\theta\theta}(0)=\frac{(F-1)^2f_0f_1f}{2(f_1F+f_0)^2}$$

and

$$S_2^2=\frac{\nabla_\theta\hat\ell(0)^2}{(-\nabla^2_{\theta\theta}\hat\ell(\hat\theta))}=\frac{(F-1)^2ff_0f_1F^2}{2(f_1F+f_0)^2}.$$

These evaluate the derivative of the profile log likelihood (4) at the null hypothesis value $\theta=0$ and rescale by either the nominal asymptotic variance at the null hypothesis $i^{\theta\theta}(0)$ or, as is sometimes recommended, by the inverse of the observed information. To impose the constraint $\theta\geq 0$, a negative gradient at $\theta=0$ may be replaced by zero, so that the test never rejects when the maximum is achieved on the boundary.

In a form comparable to (2),

$$\begin{aligned}S_1^2&=\frac{b^2\hat\theta^2f_0f_1}{2f}\Big(1-\frac{2bf_1\hat\theta}{f}\Big)+O(\hat\theta^4)\\ \Lambda/S_1^2&=1-\frac{2b\hat\theta(f_0-f_1)}{3f}-\frac{2b\hat\theta^2}{f}\Big(\frac{f(f+f_1)-6bf_1}{3f}\Big)+O(\hat\theta^3).\end{aligned}$$

Thus, the anomalous behaviour documented in sections 3.2–S.2 is not reproduced when S_1^2 is used in place of the Wald statistic. This is also visually apparent from Fig. 1, where the gradient of the profile log-likelihood function at zero is large, and the extreme non-constancy of $i^{\theta\theta}(\theta)$ does not play a role. For S_2^2 the situation is less clear, as

$$\begin{aligned}S_2^2&=\frac{b^2\hat\theta^2f_0f_1}{2f}\Big(1+\frac{2b\hat\theta f_0}{f}\Big)+O(\hat\theta^4).\\ \Lambda/S_2^2&=1-\frac{2b\hat\theta(2f+f_0)}{3f}+\frac{4b^2\hat\theta^2f_0(2f+f_0)}{3f^2}+O(\hat\theta^3),\end{aligned}$$

which consists of antagonistic expressions in $\hat\theta$ and $\hat\theta^2$, and points to anomalous behaviour if $2bf_0\hat\theta/f \gg 1$. This condition, however, is not really compatible with $\hat\theta$ being small. In terms of $\Lambda \geq 0$ the above expansions are

$$\Lambda/S_1^2 = 1 - \frac{2\sqrt{2}(f_0-f_1)\Lambda^{1/2}}{3(ff_0f_1)^{1/2}} + 4\Big(\frac{f_1f(b-3)+3f^2-2b(f_0^2-9f_1)}{9bff_0f_1}\Big)\Lambda + O(\Lambda^{3/2})$$
$$\Lambda/S_2^2 = 1 - \frac{2\sqrt{2}(2f+f_0)\Lambda^{1/2}}{3(ff_0f_1)^{1/2}} + \frac{4(2f+f_0)(6f_0+(f+f_1))\Lambda}{9ff_0f_1} + O(\Lambda^{3/2}).$$

S.4. DERIVATION OF EQUATION 10

The distribution of $W = w(Y) = U^{\mathrm{T}}Y$ is normal of mean zero and covariance $\sigma^2 A_\theta = U^{\mathrm{T}}\Sigma U$. In the transformation $w \mapsto (w/\|w\|, w^{\mathrm{T}}w) = (q, r^2)$, the volume element transforms as

$$dw_1\cdots dw_{n-p} = \frac{(r^2)^{\frac{n-p}{2}-1}dr^2\prod_{j=1}^{n-p-1}dq_j}{2\big(1-\sum_{j-1}^{n-p-1}q_j^2\big)^{1/2}} = 2^{-1}(r^2)^{\frac{n-p}{2}-1}dr^2dq$$

where $dq = dq_1\cdots dq_{n-p}$. This can be shown directly by computing the Jacobian determinant, or equivalently (Muirhead, 1984, p.50-52) by applying the exterior product calculus to the differentials of the $n-p$ equations

$$\textstyle\sum_{i=1}^j w_i^2 = r^2\sum_{i=1}^j q_i, \quad \sum_{i=1}^{n-p} q_i = 1, \quad j = 1,\dots n-p.$$

The joint density function is therefore

$$\frac{(r^2)^{(n-p)/2-1}\exp\Big(-\frac{r^2}{2\sigma^2}q^{\mathrm{T}}A_\theta^{-1}q\Big)}{2(2\pi\sigma^2)^{(n-p)/2}|A_\theta|^{1/2}}dr^2dq,$$

from which the marginal density (10) of Q is obtained by integration using a change of variables from r^2 to $r^2q^{\mathrm{T}}A_\theta^{-1}q/2\sigma^2$.

S.5. DERIVATION OF EQUATION 12

The eigenvector equation $(I_n - 1_n1_n^{\mathrm{T}}/n)u_j = u_j$ shows that U can be constructed from a Hadamard matrix of dimension n by discarding the column of ones and dividing each entry by $n^{1/2} = ((f_1+1)b)^{1/2}$.

In standardized form, in which the first row and column have all entries equal to 1, a Hadamard matrix of dimension 2^m is constructed from one of half the size as

$$H_{2^m} = \begin{pmatrix} H_{2^{m-1}} & H_{2^{m-1}} \\ H_{2^{m-1}} & -H_{2^{m-1}} \end{pmatrix}. \tag{S.6}$$

Thus, in the form most relevant for computing A_θ, U consists of repeated blocks of size $2b\times 2b$, itself consisting of sub-blocks $U_{b,1}$, $U_{b,2}$, $U_{b,3}$, $U_{b,4}$ say, obtained from the corresponding Hadamard matrices H_b by removing the first column and appending the column from the matrix immediately to the right in representation (S.6), which is either H_b or $-H_b$. These $2b\times 2b$ blocks are repeated $(f_1+1)/2$ times in the column and row dimensions, up to the $(n-1)$th column of U, which terminates an incomplete block consisting of $b\times(b-1)$ matrices $\mathring{U}_{b,2}$ and $\mathring{U}_{b,4}$, say.

Explicitly,

$$U = \begin{pmatrix} U_{b,1} & U_{b,2} & U_{b,1} & U_{b,2} & \cdots & U_{b,1} & U_{b,2} & U_{b,1} & \mathring{U}_{b,2} \\ U_{b,3} & U_{b,4} & U_{b,3} & U_{b,4} & \cdots & U_{b,3} & U_{b,4} & U_{b,3} & \mathring{U}_{b,4} \\ \vdots & \vdots & & & & \vdots & \vdots & \vdots & \vdots \\ U_{b,1} & U_{b,2} & U_{b,1} & U_{b,2} & \cdots & U_{b,1} & U_{b,2} & U_{b,1} & \mathring{U}_{b,2} \\ U_{b,3} & U_{b,4} & U_{b,3} & U_{b,4} & \cdots & U_{b,3} & U_{b,4} & U_{b,3} & \mathring{U}_{b,4} \end{pmatrix}.$$

Multiplication in blocks of size b shows that off-diagonal blocks of $U^{\mathrm{T}}(I_{f+1} \otimes 1_b 1_b^{\mathrm{T}})U$ are all zero and the sth diagonal block is

$$\begin{aligned} &(f_1+1)\big(U_{b,1}^{\mathrm{T}} 1_b 1_b^{\mathrm{T}} U_{b,1} + U_{b,3}^{\mathrm{T}} 1_b 1_b^{\mathrm{T}} U_{b,3}\big)/2 \qquad \text{s mod } 2 = 1, \\ &(f_1+1)\big(U_{b,2}^{\mathrm{T}} 1_b 1_b^{\mathrm{T}} U_{b,2} + U_{b,4}^{\mathrm{T}} 1_b 1_b^{\mathrm{T}} U_{b,4}\big)/2 \qquad \text{s mod } 2 = 0, \end{aligned}$$

the terminal diagonal block being

$$(f_1+1)\big(\mathring{U}_{b,2}^{\mathrm{T}} 1_b 1_b^{\mathrm{T}} \mathring{U}_{b,2} + \mathring{U}_{b,4}^{\mathrm{T}} 1_b 1_b^{\mathrm{T}} \mathring{U}_{b,4}\big)/2$$

All columns of $U_{b,j}$ except the last sum to zero for $j = 1, \ldots, 4$, the bth column summing to $\pm b/((f_1+1)b)^{1/2}$. Thus, in

$$A_\theta = I_{n-1} + \theta U^{\mathrm{T}}(I_{f+1} \otimes 1_b 1_b^{\mathrm{T}})U,$$

$U^{\mathrm{T}}(I_{f+1} \otimes 1_b 1_b^{\mathrm{T}})U$ is a matrix of zeros with entry b at diagonal positions $b, 2b, \ldots, f_1 b$. It follows that $\log|A_\theta| = f_1 \log(1+b\theta)$ and

$$q^{\mathrm{T}} A_\theta^{-1} q = 1 + \textstyle\sum_{j=1}^{f_1} q_{jb}^2 \big((1+b\theta)^{-1} - 1\big).$$

Resolve $v = UU^{\mathrm{T}} y$ into orthogonal components $v = v' + v''$, where $v' \in V \subset \mathbb{R}^n$ and $v'' \in V^\perp$. More explicitly, the components of v are

$$v_{jb} = \overline{v}_{\bullet\bullet} + (\overline{v}_{j\bullet} - \overline{v}_{\bullet\bullet}) + (v_{jb} - \overline{v}_{j\bullet}) = (\overline{v}_{j\bullet} - \overline{v}_{\bullet\bullet}) + (v_{jb} - \overline{v}_{j\bullet}).$$

For any fixed $s \in \{1, \ldots, b\}$, $\{u_{\bullet s}, u_{\bullet(2s)}, \cdots, u_{\bullet(f_1 s)}\}$ is an orthonormal basis for the $(k-1)$-dimensional subspace V, where $u_{\bullet s}$ denotes the sth column of U, the remaining $k(b-1)$ columns of U being a basis for $V^\perp$. It follows that

$$v' = \sum_{j=1}^{f_1} \langle v - v'', u_{\bullet(jb)} \rangle u_{\bullet(jb)} = \sum_{j=1}^{f_1} \langle y, u_{\bullet(jb)} \rangle u_{\bullet(jb)} = \sum_{j=1}^{f_1} w_{jb} u_{\bullet(jb)}$$

and by orthonormality of the basis, $\|v'\|^2 = \sum_{j=1}^{f_1} w_{jb}^2$ so that

$$\sum_{j=1}^{f_1} q_{jb}^2 = \frac{f_1 \mathrm{MS}_1}{f_0 \mathrm{MS}_0 + f_1 \mathrm{MS}_1} = \frac{f_1 F}{f_0 + f_1 F}.$$

REFERENCES

[1] MUIRHEAD, R. J. (1984). *Aspects of Multivariate Statistical Theory*, Wiley, Hoboken.

[*Received on 26 January 2023. Editorial decision on 00* 000 *0000*]