


Improved inference for a boundary parameter

Soumaya ELKANTASSI^{1*}, Ruggero BELLIO², Alessandra R. BRAZZALE³,
 and Anthony C. DAVISON⁴ 

¹Department of Operations, University of Lausanne, Lausanne, Switzerland

²Department of Economics and Statistics, University of Udine, Udine, Italy

³Department of Statistical Sciences, University of Padova, Padova, Italy

⁴Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Key words and phrases: Higher order likelihood inference; mixture model; nonstandard likelihood theory; restricted likelihood; smoothing spline.

MSC 2020: Primary 62E20; secondary 62F05.

Abstract: The limiting distributions of statistics used to test hypotheses about parameters on the boundary of their domains may provide very poor approximations to the finite-sample behaviour of these statistics, even for very large samples. We review theoretical work on this problem, describe hard and soft boundaries and iceberg estimators, and give examples highlighting how the limiting results greatly underestimate the probability that the parameter lies on its boundary even in very large samples. We propose and evaluate some simple remedies for this difficulty based on normal approximation for the profile score function, and then outline how higher order approximations yield excellent results in a range of hard and soft boundary examples. We use the approach to develop an accurate test for the need for a spline component in a linear mixed model.

Résumé: Les statistiques utilisées pour tester des hypothèses concernant des paramètres situés sur la frontière de leur domaine présentent souvent des distributions limites non standard, qui peuvent être de mauvaises approximations d'échantillon fini même lorsque la taille de l'échantillon est très grande. Nous passons en revue les travaux théoriques relatifs à ce problème, décrivons les problèmes de limites dures et molles et les estimateurs "iceberg", et donnons des exemples montrant que les résultats limites sous-estiment fortement la probabilité que le paramètre se trouve sur sa frontière, même lorsque l'échantillon est d'une taille importante. Nous proposons et évaluons quelques remèdes simples basés sur l'approximation normale de la fonction score profilée; puis nous démontrons que les approximations d'ordre supérieur donnent d'excellents résultats dans une série d'exemples de frontières dures et molles. Nous utilisons l'approche pour développer des tests pour une composante spline dans un modèle linéaire mixte.

Dedication

Nancy Reid has played a leading role in developments in statistical theory and in advancing the discipline of statistics by stimulating and supporting research in Canada and more widely. She has also been a wise and wonderful mentor and friend, to whom we offer this small birthday present as a token of our gratitude, respect, and admiration.

* Corresponding author: Soumaya.Elkantassi@unil.ch

1. INTRODUCTION

Inference procedures based on likelihood theory form the backbone of many statistical methods, owing to the appeal of likelihood as a measure of plausibility, the generality of the likelihood paradigm, the flexibility with which new problems can be addressed, and close links to the Bayesian approach. The original notion now encompasses a wide range of related ideas, including conditional and marginal likelihoods, partial likelihoods, empirical likelihoods, quasi- and pseudo-likelihoods, and composite likelihoods; see, for example, Pawitan (2001). An appealing aspect is that the standard theory, which was initiated by Fisher (1922, 1925, 1934), leads to a few well-understood, simple, and widely applicable approximations for inference. These typically rely on normal and chi-squared distributions and have been easily applied from well before the computer age. Over the past few decades, this classical theory has undergone substantial further development in which Nancy Reid has played a central role, and in its modern form it can yield highly accurate inferences based on parametric models even for very small samples (Barndorff-Nielsen & Cox, 1994; Fraser & Reid, 1995; Severini, 2000; Brazzale, Davison & Reid, 2007; Fraser, 2017).

Much less attention has been paid to so-called nonregular cases, under which the standard conditions for validity of these classical approximations do not hold. These conditions, which are typically of the Cramér type (Cramér, 1946, §33.3), include differentiability of the underlying joint probability or density function up to a suitable order and finiteness of the Fisher information matrix, but fail for models that are commonly used in genetics, econometrics, and many other fields. One class of examples consists of the so-called endpoint problems, in which the support of the observations must be estimated; in this case, the shape of the density function at the limits of its support determines the accuracy with which the endpoint, and possibly other parameters, can be estimated (Smith, 1985). Nonregularity can arise in many other ways; see Smith (1989), Cheng & Traylor (1995), Barndorff-Nielsen & Cox (1994, §3.8), Davison (2003, §4.6), and Cox (2006, Chapter 7).

Brazzale & Mameli (2023) group nonregular problems into three broad classes. One class consists of change-point problems, and a related class is situations in which one or more components of the parameter vanish when another component is set to a particular value. In both cases, approximate distributions for likelihood-based statistics can be complex and their usefulness may be limited in practice. A third class of nonregular problems comprises the so-called boundary cases, where it is desired to test the hypothesis that some interest parameter ψ equals a null value ψ_0 against the alternative that $\psi > \psi_0$, and ψ_0 lies on the boundary of its domain. Informally, the resulting difficulties occur because the maximum likelihood estimator can fall only on one side of ψ_0 . If the maximum occurs on the boundary ψ_0 , the score function need not be zero there, and the distributions of the related likelihood statistics will not converge to the typical normal or chi-squared limits. Because of the difficulties in deriving and then in using the appropriate limiting distributions, practitioners tend to ignore the boundary problem and to apply the usual asymptotic results. This naïve approach can lead to very inaccurate inferences, especially for complex models.

In this article, we study finite-sample approximations for certain boundary problems and show how they may be greatly improved using higher order likelihood procedures. The only precursor paper of which we are aware is by Castillo & López-Ratera (2006), who demonstrate the validity of an improved signed likelihood ratio statistic when testing a boundary hypothesis on a scalar parameter in an exponential family. A large literature on higher order likelihood inference for regular models, in both classical and Bayesian frameworks, summarized in Brazzale, Davison & Reid (2007), Severini (2000), or Barndorff-Nielsen & Cox (1994), shows how highly accurate approximations to the distributions of test statistics and pivots may be obtained for a variety of parametric statistical models.

We distinguish between soft and hard boundaries. An example of the first, testing in simple mixed effects models, includes as a special case comparison of parametric regression models with semiparametric alternatives, under the now-standard formulation of spline regression as a linear mixed model (Laird & Ware, 1982; McCulloch & Searle, 2001; Ruppert, Wand & Carroll, 2003; Wood, 2017). This is a soft boundary because, although the parameter to be tested makes *statistical* sense only when it is nonnegative, the likelihood is *mathematically* well defined in an open neighbourhood around the null value, so standard higher order approximations can be applied on the boundary. An example of a hard boundary arises in testing for infinite mixtures, such as the Student- t distribution with continuous degrees of freedom $\nu = \psi^{-1}$ or the negative binomial distribution with $k = \psi^{-1}$ successes, which reduce to the normal and Poisson models, respectively, when $\psi = 0$. This is a hard boundary because ψ cannot be negative.

The asymptotic distribution of the likelihood ratio statistic in both settings is typically $\bar{\chi}^2$, a finite mixture of chi-squared variables with known probabilities and degrees of freedom (Self & Liang, 1987), although other limiting distributions can also arise (Sinha, Kopylev & Fox, 2012). In the simplest case, the limiting distribution of the likelihood ratio statistic puts mass 1/2 at $\psi = 0$, with the remaining probability spread as a χ_1^2 distribution, although we shall see that this approximation provides unreliable inferences even in large samples.

The rest of the article is organized as follows: Section 2 briefly reviews the literature on boundary problems and motivates our work. Section 3 describes some simple remedies that use corrections to the profile score to better estimate the mass on the boundary, and explores their adequacy. Section 4 reviews higher order likelihood theory and illustrates how well it can be used to improve the entire distribution of likelihood-based test statistics in both soft and hard boundary problems. Section 5 describes and illustrates an application of the ideas to the linear mixed model. Our conclusions are given in Section 6. Full details of the calculations and further numerical results may be found in Chapter 3 of Elkantassi (2023).

2. BOUNDARY PROBLEMS

2.1. Background

Let $\ell(\psi, \lambda)$ denote the log-likelihood function for a parametric statistical model for data with sample size n , with scalar parameter of interest ψ and nuisance parameter λ , where $\psi \in \Psi = [\psi', \infty)$ for some known real number ψ' and the value of λ that generated the data is interior to an open set Λ . Suppose we wish to test the null hypothesis $H_0 : \psi = \psi_0$ for some fixed known ψ_0 . If $\psi_0 > \psi'$, then under mild regularity conditions and in large samples, the signed likelihood ratio statistic

$$r(\psi_0) = \text{sign}(\hat{\psi} - \psi_0)[2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi_0, \hat{\lambda}_0)\}]^{1/2}, \quad (1)$$

also known as the likelihood root, has a standard normal distribution, and the likelihood ratio statistic $w(\psi_0) = r(\psi_0)^2$ follows a χ_1^2 distribution. In Equation (1), $(\hat{\psi}, \hat{\lambda})$ denotes the overall maximum likelihood estimators of the parameters, and $\hat{\lambda}_0$ denotes the maximum likelihood estimator of λ when $\psi = \psi_0$. If $\psi_0 = \psi'$, then a boundary problem occurs.

Research on boundary problems was initiated by Chernoff (1954), who derived the asymptotic null distribution of the likelihood ratio statistic for testing where ψ lies relative to a smooth $(p - 1)$ -dimensional surface in a p -dimensional space when the true parameter value lies on the surface. Geometrical arguments establish that this distribution is equivalent to that of the likelihood ratio statistic for testing restrictions on the mean of a multivariate normal distribution with a known covariance matrix. An important further step is by Self & Liang (1987), who allow ψ_0 to lie on the boundary of the parameter space and suppose that the latter could be approximated by a tangent cone with vertex at ψ_0 ; Geyer (1994) discusses these limiting results

further. Kopylev & Sinha (2011) and Sinha, Kopylev & Fox (2012) derive the null distribution of the likelihood ratio statistic through algebraic arguments. The difficulty of deriving closed-form expressions for such distributions increases when more nuisance parameters lie on the boundary of the parameter space. All these contributions are summarized by Kopylev (2012), who also describes applications in genetics and biology. In general terms, the asymptotic cumulative distribution is chi-bar squared (Kudo, 1963):

$$\Pr(\bar{\chi}^2 \leq x) = \sum_{v=0}^N \omega_v \Pr(\chi_v^2 \leq x) \quad \text{for } x \in \mathbb{R}, \quad (2)$$

i.e., a mixture of chi-squared distributions with degrees of freedom v and nonnegative weights ω_v summing to unity that depend on the number and type of parameters and on the geometry of the tangent cone; the χ_0^2 distribution places a point mass on $x = 0$. In the following, we compare the finite-sample probability that the maximum likelihood estimator is positive to the asymptotic weight $\omega_1 = 1/2$ in the simplest situation, i.e., $N = 1$.

2.2. Smoking Guns

The asymptotic approximation given by Equation (2) can perform very poorly. Consider the one-way random effects model

$$Y_{ij} = \mu + b_i + \varepsilon_{ij} \quad \text{for } i = 1, \dots, k \text{ and } j = 1, \dots, m, \quad (3)$$

where μ is the overall mean, and b_i and ε_{ij} are mutually independent normal random variates having zero means and respective variances σ_b^2 and σ^2 . This corresponds to a sample of independent observations divided into k groups, each of size m , so $n = mk$. Let $\psi = \sigma_b^2/\sigma^2$ and set $\lambda = (\mu, \sigma^2)$, so that a test of the boundary hypothesis $H_0 : \psi = \psi_0 = 0$ corresponds to testing $b_1 = \dots = b_k = 0$. In this case, an exact F test is available, but it is instructive to apply the large-sample approximation. The large-sample distribution of the likelihood root is (Chernoff, 1956)

$$pr_0\{r(\psi_0) \leq x\} = I(x \geq 0)\Phi(x) \quad \text{for } x \in \mathbb{R}, \quad (4)$$

where pr_0 denotes the probability under H_0 , $I(\cdot)$ is an indicator function, and $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable. The P -value for a test of H_0 is thus $p_{\text{obs}} = pr_0\{r(\psi_0) \geq r_{\text{obs}}\}$, where r_{obs} is the observed value of $r(\psi_0)$, and $pr_0(\cdot)$ denotes a probability computed under H_0 . If $\hat{\psi} = \psi_0$, i.e., the maximum likelihood estimate of ψ lies on the boundary, then $r_{\text{obs}} = 0$ and Equation (4) yields $p_{\text{obs}} = 1$, whereas if $r_{\text{obs}} > 0$, i.e., $\hat{\psi} > \psi_0$, then Equation (4) yields $p_{\text{obs}} = \Phi(-r_{\text{obs}})$.

Apart from additive constants, the log-likelihood function may be written as

$$\begin{aligned} \ell(\mu, \sigma^2, \psi) = & -\frac{1}{2} \left\{ mk \log \sigma^2 + \frac{C_2}{\sigma^2} + \frac{C_1}{\sigma^2(1+m\psi)} \right. \\ & \left. + k \log(1+m\psi) + \frac{km(\bar{y}_{..} - \mu)^2}{\sigma^2(1+m\psi)} \right\}, \end{aligned} \quad (5)$$

where the grand mean $\bar{y}_{..}$ and the sums of squares C_1 and C_2 between groups and within groups are mutually independent and satisfy

TABLE 1: Probability (%) of positive maximum-likelihood estimator of the variance ratio in model (3).

k	$m = 5$		$m = 10$		$m = 20$		$m = 30$	
	Usual	REML	Usual	REML	Usual	REML	Usual	REML
5	32.2	43.0	30.3	41.7	29.5	41.1	29.2	40.9
10	37.7	45.5	36.3	44.6	35.6	44.1	35.4	44.0
20	41.4	47.0	40.4	46.3	39.9	45.9	39.7	45.8
30	43.0	47.6	42.1	47.0	41.7	46.7	41.6	46.6
50	44.6	48.1	43.9	47.7	43.6	47.5	43.5	47.4
100	46.2	48.7	45.7	48.4	45.5	48.2	45.4	48.2

Abbreviation: REML, restricted maximum likelihood.

$$C_1 = m \sum_{i=1}^k (\bar{y}_i - \bar{y}_{..})^2 \sim \sigma^2(1 + m\psi) \chi_{k-1}^2,$$

$$C_2 = \sum_{i=1}^k \sum_{j=1}^m (y_{ij} - \bar{y}_{i.})^2 \sim \sigma^2 \chi_{k(m-1)}^2$$

with $\bar{y}_{..}$ and $\bar{y}_{i.}$, respectively, representing the overall average and that for the i th group. It is easily checked that the profile log-likelihood function for ψ may be written

$$\ell_p(\psi) \equiv -\frac{k}{2} \left\{ m \log \left(C_2 + \frac{C_1}{1 + m\psi} \right) + \log(1 + m\psi) \right\} \quad \text{for } \psi \geq 0, \quad (6)$$

and using the likelihood root given in Equation (1), we have $pr_0 \{r(\psi_0) > 0\} = pr\{B > m^{-1}\}$, where the random variable B has the beta distribution with parameters $(k-1)/2$ and $k(m-1)/2$. Table 1 shows that this probability is markedly smaller than 0.5 for any values of k and m likely to arise in practice; the limiting value appears as $k \rightarrow \infty$, and the asymptotic approximation degrades slightly as m increases for fixed k .

Likelihood-based variance estimation is downwardly biased, which suggests using the restricted log-likelihood function (Harville, 1977). This amounts to dropping the last term in Equation (5) and replacing the coefficients mk and k for the logarithmic terms by $mk-1$ and $k-1$, respectively. The resulting probability of a positive gradient, $pr\{B > (k-1)/(km-1)\}$, is also given in Table 1. The approximation is better, but remains too poor for reliable use in applications.

The results for the previous example might be ascribed to the estimation of a variance. To see that similar problems can arise in other settings, consider a random sample of size n from a mixture of p -dimensional normal variables, $\frac{1}{2} \mathcal{N}_p(\lambda - \psi \mathbf{1}_p, I_p) + \frac{1}{2} \mathcal{N}_p(\lambda + \psi \mathbf{1}_p, I_p)$, where $\lambda \in \mathbb{R}^p$ is a vector of nuisance parameters, $\psi \geq 0$ is a scalar interest parameter, and $\mathbf{1}_p$ and I_p are, respectively, a $p \times 1$ vector of 1 and the $p \times p$ identity matrix. The parameters λ and ψ can be estimated using the EM algorithm, and when $p = 2$ and $n = 30, 50, 100, 200,$ and 500 , the probabilities (%) that $\hat{\psi} > 0$ are, respectively, 41.2, 44.3, 45.2, 46.8, and 48.9. Increasing p up to 20 does not much affect the results. Although ψ is a location parameter and the variances are known, the probability of a boundary case, $\hat{\psi} = 0$, is appreciably larger than the asymptotic results suggest.

In these and other examples, we have considered the empirical probability of boundary cases in finite samples typically exceeding its asymptotic value, often by some margin. In the next section we describe some simple fixes for this problem.

3. SIMPLE SOLUTIONS

3.1. Icebergs

Feng & McCulloch (1992) suggest that in some cases one can avoid difficulties with boundaries and improve on the naïve approach by extending the domain of the log-likelihood function to all parameter values for which the density function is mathematically well defined, even if certain values make no sense statistically, while introducing a rule to impose the usual large-sample distributions for likelihood-based statistics. Their method is illustrated on a three-component mixture of normals with one or two mixing proportions on their boundary, in which setting the true coverage probabilities of confidence regions constructed by their “intersection method” are acceptable; however, this approach works only when the null hypothesis is uniquely identified. A counterexample is given by Dietz & Böhning (1995) in their discussion of Cheng & Traylor (1995).

This idea suggests that one might visualize a maximum likelihood estimator $\tilde{\psi}$, which equals $\hat{\psi}$ when the latter is positive but can take also values outside the parameter space. In the simplest case, $\tilde{\psi} \sim \mathcal{N}(\delta, \tau^2)$ and $\hat{\psi} = \max(\tilde{\psi}, 0)$. We call this an iceberg because the part of its distribution outside the parameter space is invisible, and it is typically larger than the visible part. In analogy to Equation (1), we can equally envisage an iceberg likelihood root, $\tilde{r}(\psi_0) \doteq \mathcal{N}(\delta/\tau, 1)$. When $\delta < 0$, then $pr\{\hat{\psi} > 0\} = pr\{\tilde{\psi} > 0\} \doteq \Phi(\delta/\tau) < 1/2$, and the corresponding P -value $\Phi(-r_{\text{obs}})$ computed using Equation (4) will be too large, leading to a loss of power when testing the boundary hypothesis using the large-sample results. For a crude remedy, we then write $pr\{\tilde{r}(\psi_0) > 0\} \doteq p_+$, giving $\delta/\tau \doteq \Phi^{-1}(p_+)$ and the improved P -value

$$\begin{aligned} pr_0\{r(\psi_0) > r_{\text{obs}}\} &= pr_0\{\tilde{r}(\psi_0) > r_{\text{obs}}\} \\ &\doteq 1 - \Phi(r_{\text{obs}} - \delta/\tau) \\ &\doteq \Phi\{\Phi^{-1}(p_+) - r_{\text{obs}}\} \quad \text{for } r_{\text{obs}} > 0. \end{aligned} \quad (7)$$

Here and elsewhere, $\Phi^{-1}(\cdot)$ denotes the inverse of the function $\Phi(\cdot)$. This argument presupposes that $\tilde{\psi}$ has an approximate normal distribution and can only be applied when p_+ can be computed or adequately approximated. We now discuss one approach to doing so, by obtaining analytical estimates of p_+ .

3.2. Profile Score Modification

Under mild conditions, the log-likelihood function will have a local maximum on the boundary if the profile log-likelihood function for ψ

$$\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi) = \max_\lambda \ell(\psi, \lambda),$$

has a negative gradient there, i.e., $\hat{\psi} > \psi_0$ if $\partial \ell_p(\psi)/\partial \psi > 0$ at $\psi = \psi_0$. This suggests considering

$$p_+ = pr_0(\hat{\psi} > \psi_0) \doteq pr_0 \left\{ \left. \frac{\partial \ell_p(\psi)}{\partial \psi} \right|_{\psi=\psi_0} > 0 \right\}, \quad (8)$$

where pr_0 denotes the probability computed under the model with $\psi = \psi_0$, and the approximation arises because the global maximum of the likelihood might lie away from the boundary even

if the profile score has a negative derivative there. Unlike the distribution of the maximum likelihood estimator, that of the profile score statistic is continuous, without a point mass at the origin. In an investigation of its finite-sample properties for a random sample of n observations, McCullagh & Tibshirani (1990) show that

$$\begin{aligned} E \left\{ \frac{\partial \ell_p(\psi)}{\partial \psi} \right\} = & -\frac{1}{2} \left\{ (\kappa_{\psi,r,s} - \kappa_{\psi,t} \kappa^{t,u} \kappa_{u,r,s}) \kappa^{r,s} \right. \\ & \left. + (\kappa_{\psi,r,s} - \kappa_{\psi,t} \kappa^{t,u} \kappa_{u,r,s}) \kappa^{r,s} \right\} + O(n^{-1/2}), \end{aligned} \quad (9)$$

where summation is implied over repeated indices. The indices r, s, t, u refer to components of the nuisance parameter λ ; $\kappa_{r,s}$ is the (r, s) component of the Fisher information matrix for λ ; $\kappa^{r,s}$ is a component of the corresponding inverse matrix; $\kappa_{\psi,r}$ is a component of the Fisher information corresponding to ψ and λ_r ; and

$$\kappa_{\psi,r,s} = E \left\{ \frac{\partial \ell}{\partial \psi} \frac{\partial \ell}{\partial \lambda_r} \frac{\partial \ell}{\partial \lambda_s} \right\}, \quad \kappa_{\psi,r,s} = E \left\{ \frac{\partial \ell}{\partial \psi} \frac{\partial^2 \ell}{\partial \lambda_r \partial \lambda_s} \right\}.$$

The terms $\kappa_{r,s}$, $\kappa_{\psi,r,s}$, and so on, are $O(n)$, and $\kappa^{r,s}$ is $O(n^{-1})$, so the leading term on the right-hand side of Equation (9) is $O(1)$. A similar computation gives the familiar formula

$$\text{var} \left\{ \frac{\partial \ell_p(\psi)}{\partial \psi} \right\} = \kappa_{\psi,\psi} - \kappa_{\psi,r} \kappa_{s,\psi} \kappa^{r,s} + O(n^{1/2}), \quad (10)$$

whose leading term is $O(n)$. If we call the leading terms on the right-hand sides of Equations (9) and (10) m_p and s_p^2 , respectively, then a normal approximation gives

$$p_+ \doteq \Phi(m_p/s_p) = p_+^a.$$

If it is feared that such an approximation is inadequate, then refinement might be sought using the first term of an Edgeworth expansion, which requires the skewness k_p of the profile score. Painful computations sketched in Elkantassi (2023) lead to an approximation to k_p , enabling a more refined version of p_+^a , which we label p_+^e . We shall see later that this does not help much, so we omit the details.

3.3. Numerical Examples

We assess the value of the preceding computations by comparing the generalized Pareto, Student- t , and negative binomial distributions with baseline exponential, normal, and Poisson distributions, respectively. The first three can be generated as infinite mixtures of the last three, but the first has the peculiarity that although the mixture results in overdispersion relative to the corresponding baseline, the generalized Pareto can also be less dispersed than its baseline, giving a soft boundary, whereas the other two examples have hard boundaries. We now give some more details.

A generalized Pareto variable with positive shape parameter $1/\nu$ can be expressed as X/Z , where X is standard exponential and Z is gamma with mean $1/\sigma$ and variance $1/(\nu\sigma^2)$, and X and Z are independent. The exponential distribution appears as $\nu \rightarrow \infty$, so testing $\psi = 1/\nu = 0$ against $\psi > 0$ corresponds to testing that the infinite mixture of exponential distributions generated by averaging over Z is degenerate, because $Z = 1/\sigma$ with probability 1 when $\psi = 0$. Taking $\psi < 0$ also yields a valid distribution, so the boundary $\psi = 0$ is soft. The moments of the profile score for ψ at $\psi = 0$ are $m_p = -1$, $s_p^2 = n$, and $k_p = 13n$.

TABLE 2: Estimated probabilities (%) of positive maximum likelihood estimators when testing $\psi = 0$ in generalized Pareto, Student- t , and negative binomial models for various sample sizes n .

		n				
		10	20	50	100	200
Generalized Pareto	p_+	39.2 ^a	32.9	37.2	40.5	43.1
	p_+^a	37.6	41.1	44.4	46.0	47.2
	p_+^e	36.9	41.0	44.3	46.0	47.2
Student- t	p_+	21.1	26.3	33.5	36.3	39.7
	p_+^a	34.9	39.2	43.1	45.1	46.5
	p_+^e	33.6	38.8	43.0	45.1	46.5
Negative binomial	p_+	33.5	37.1	41.5	44.3	45.9
	p_+^a	41.1	43.7	46.0	47.2	48.0
	p_+^e	40.9	43.6	46.0	47.2	48.0

Note: The probabilities p_+ , p_+^a , and p_+^e are, respectively, obtained using 10^4 simulated samples and profile score adjustment, without and with Edgeworth expansion.

^aThis figure is unreliable owing to maximization problems in very small samples.

A Student- t variable centred at μ and with dispersion parameter σ^2 and degrees of freedom ν can be expressed as $\mu + \sigma Z / \sqrt{V/\nu}$, where Z is standard normal, V has the chi-squared distribution with ν degrees of freedom, and Z and V are independent. Thus, setting $\psi = 1/\nu$ and testing for $\psi = 0$ amounts to testing for normality against the infinite mixture arising when $\psi > 0$. In this case, the moments of the profile score are $m_p = -3/2$, $s_p^2 = 3n/2$, and $k_p = 369n/8$.

A negative binomial variable can be constructed by taking a Poisson variable with mean λX , where X is gamma with unit mean and shape parameter ν . Setting $\psi = 1/\nu$ and testing for $\psi = 0$ against $\psi > 0$ corresponds to testing the baseline Poisson distribution. Here, $m_p = -\lambda/2$, $s_p^2 = n\lambda^2/2$, and $k_p = n\lambda^2(\lambda + 2)/4$.

Table 2 compares the empirical probabilities p_+ that $\hat{\psi} > 0$ in each of these examples with estimates p_+^a and p_+^e obtained using the moments of the profile score. The values of p_+ can be much less than 0.5 even when $n = 200$. Those of p_+^a lie roughly half-way between p_+ and 0.5, so although they offer improvements, they fail to track p_+ closely. The values of p_+^e improve slightly on those of p_+^a but not by enough to justify computing the third cumulant.

Although not detailed here, calculations for the variance components model specified in Equation (3) lead to similar conclusions: corrections to the profile score such as these do not provide reliable inferences. In the next section, we study a different approach applicable for soft boundaries, whereby we aim to improve the limiting normal distributions of boundary test statistics.

4. HIGHER ORDER APPROXIMATION

4.1. Modified Likelihood Root

Classical large-sample likelihood approximations of the type described in the previous sections have long been known to have drawbacks, especially in the presence of high-dimensional nuisance parameters, and intensive work over the past four decades has led to major developments in which Nancy Reid and Don Fraser have played a crucial role. Although there are numerous closely

related approaches, a relatively simple modification to the first-order methods in regular settings is the replacement of the likelihood root defined in Equation (1) by the modified likelihood root (Barndorff-Nielsen, 1986):

$$r^*(\psi_0) = r(\psi_0) + \frac{1}{r(\psi_0)} \log \left\{ \frac{q(\psi_0)}{r(\psi_0)} \right\}. \quad (11)$$

This expression combines the likelihood root $r(\psi)$ with a correction $q(\psi)$ that incorporates both an adjustment for the elimination of the nuisance parameters and for any departure from normality of $r(\psi)$. Detailed discussions of higher order approximation and expressions for the correction term q in Equation (11) may be found in Fraser, Reid & Wu (1999), Reid (2003), and Brazzale et al. (2007, Chapter 8).

The distribution of the likelihood root $r(\psi)$ based on continuous responses is standard normal up to first order, i.e., the true and nominal coverages of one-sided confidence intervals differ asymptotically by $\mathcal{O}(n^{-1/2})$, whereas $r^*(\psi_0)$ is standard normal up to third order, implying that this difference reduces to $\mathcal{O}(n^{-3/2})$; it is $\mathcal{O}(n^{-1})$ for discrete responses. Perhaps more important in applications, the resulting error is relative in large deviation regions, implying that the accuracy is retained far into the distributional tail.

The new ingredient required to apply Equation (11) is the correction $q(\psi)$, a fairly general form of which can be based on a local exponential family approximation to the data density, known as a tangent exponential model (Fraser & Reid, 1993, 1995). When the sample consists of independent continuous observations that are informative about a vector parameter $\theta = (\psi, \lambda)$ of dimension d , the local canonical parameter of this approximate exponential family can be written as

$$\varphi^\top(\theta) = \sum_{j=1}^n V_j \frac{\partial \ell(\theta; y)}{\partial y_j} \Bigg|_{y_j=y_j^o}, \quad (12)$$

where $y^o = (y_1^o, \dots, y_n^o)$ denotes the realized value of the observation vector $y = (y_1, \dots, y_n)$. The right-hand side of Equation (12) is the derivative of $\ell(\theta)$ in the directions of the d columns of the $n \times d$ matrix V . In the continuous setting and with independent data, this matrix can be constructed using a vector of pivotal quantities $z = (z_1(y_1, \theta), \dots, z_n(y_n, \theta))$, as

$$V = - \left\{ \left(\frac{\partial z^\top}{\partial y} \right)^{-1} \left(\frac{\partial z}{\partial \theta^\top} \right) \right\} \Bigg|_{(y^o, \hat{\theta}^o)}, \quad (13)$$

where $\hat{\theta}^o$ is the maximum likelihood estimate corresponding to y^o . An expository account of this construction, including further details and a brief history of its development, is given by Davison & Reid (2023).

The correction term appearing in Equation (11) can be written in terms of the local canonical parameter of the tangent exponential model as

$$q(\psi) = \frac{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi) \quad \varphi_\lambda(\hat{\theta}_\psi)|}{|\varphi_\theta(\hat{\theta})|} \times \left\{ \frac{|J(\hat{\theta})|}{|J_{\lambda\lambda}(\hat{\theta}_\psi)|} \right\}^{1/2},$$

where $\varphi_\theta(\theta)$ and $\varphi_\lambda(\theta)$ denote the $d \times d$ and $d \times (d-1)$ matrices of partial derivatives $\partial \varphi / \partial \theta^\top$ and $\partial \varphi / \partial \lambda^\top$, and the difference $\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi)$ appearing in the first numerator determinant is a

TABLE 3: Empirical probability (%) of the positive maximum likelihood estimator of the variance ratio in (3) based on 10^4 simulated datasets using the likelihood root r and modified likelihood root r^* for the standard and restricted likelihoods, for a subset of the cases in Table 1.

k	pivot	$m = 5$		$m = 10$		$m = 20$		$m = 30$	
		Usual	REML	Usual	REML	Usual	REML	Usual	REML
5	r	32.8	43.5	29.8	41.1	30.2	41.2	29.6	41.3
	r^*	48.7	50.3	47.6	49.0	48.5	49.7	49.2	50.5
10	r	37.1	44.9	36.2	44.5	36.7	45.0	35.6	44.0
	r^*	49.3	49.7	49.5	49.8	50.3	50.6	49.5	49.9
20	r	40.6	46.1	39.9	45.8	39.9	46.1	40.2	46.5
	r^*	49.0	49.1	49.3	49.5	49.7	49.9	50.5	50.6

Note: Bold figures are within simulation error of 50%

TABLE 4: Empirical probability (%) of positive maximum likelihood estimator of $\psi = 1/\nu$ for generalized Pareto samples, based on 10^4 simulated datasets of size n .

n	40	60	80	100	200	400
r	36.0	38.3	40.5	40.5	43.1	46.0
r^*	50.0	50.2	50.8	50.4	49.7	50.8

Note: Bold figures are within simulation error of 50%

$d \times 1$ vector. The second term on the right-hand side involves the $d \times d$ observed information matrix at the overall maximum likelihood estimate, $j(\hat{\theta})$, and its $(d - 1) \times (d - 1)$ submatrix $J_{\lambda\lambda}(\hat{\theta}_\psi)$ corresponding to λ evaluated at $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$. These expressions may appear complex, but the only additional quantities needed beyond the values of the log-likelihood function and its second derivative evaluated at the overall and constrained maximum likelihood estimates $\hat{\theta}^0$ and $\hat{\theta}_\psi^0$, which are needed for first-order inference, are $\varphi(\theta)$ and its derivatives, and these are typically straightforward to compute.

4.2. Numerical Examples

The ideas sketched above require classical regularity conditions for likelihood inference, but can also be applied in certain soft boundary problems. The profile log-likelihood function specified in Equation (6) for the variance ratio $\psi = \sigma_b^2/\sigma^2$ from the model given in Equation (3) is mathematically well defined when $\psi > -1/m$, so the higher order approach can be used to test the boundary hypothesis $H_0 : \psi = 0$. Table 3, which illustrates how this approach performs for a subset of Table 1, shows that r^* provides essentially correct inferences in almost all cases, with a slight improvement if it is applied to the residual likelihood. Quantile–quantile plots for simulated data show that the distribution of r^* is almost exactly standard normal, so that tests of H_0 are essentially perfectly calibrated. Similar results are obtained in a two-way layout.

As mentioned in Section 3.3, the boundary hypothesis $\psi = 0$ for testing an exponential distribution against a more dispersed generalized Pareto alternative is also soft. Table 4 and quantile–quantile plots show that higher order methods work essentially perfectly in this case also.

TABLE 5: Empirical coverages (%) for upper confidence limits α based on 10^4 simulations from the Student- t distribution with reciprocal degrees of freedom $\psi = n^{-(1+\epsilon)/2}$, for sample sizes $n = 20, 50,$ and 100 and $\epsilon = 0.2, 0.4,$ and 1 .

		$n = 20$		$n = 50$		$n = 100$	
α (%)		r	r^*	r	r^*	r	r^*
$\epsilon = 0.2$	0.5	0.2	0.4	0.3	0.5	0.2	0.4
	5	3.1	5.0	3.3	5.0	3.2	4.9
	10	6.4	10.3	6.7	10.1	6.7	9.9
	20	12.9	20.2	14.0	20.4	14.7	20.0
	30	19.4	29.9	21.2	29.8	22.2	30.3
$\epsilon = 0.4$	40	26.2	30.4	28.8	33.8	30.6	35.2
	0.5	0.3	0.6	0.2	0.5	0.3	0.7
	5	3.2	5.1	2.9	4.8	3.3	5.1
	10	6.0	10.0	6.1	10.1	6.8	10.2
	20	12.1	20.8	12.7	19.5	14.2	21.3
$\epsilon = 1$	30	18.2	28.6	19.6	30.3	21.9	31.3
	40	25.1	28.6	27.4	32.3	29.8	33.9
	0.5	0.3	0.5	0.3	0.6	0.4	0.9
	5	2.5	4.9	2.6	5.3	3.0	5.4
	10	5.0	10.5	5.6	10.4	6.0	11.3
	20	11.0	22.1	12.0	22.0	13.3	22.5
	30	16.9	26.0	19.2	28.8	20.1	32.0
	40	22.6	26.2	26.0	29.9	28.3	32.9

Note: Figures in bold are within simulation error of the nominal values.

Hard boundaries are more difficult to deal with, because obtaining the derivatives required for the higher order approximations may be tricky near the boundary, and singular observed information matrices and other headaches may arise. This is not the case for the negative binomial example of Section 3.3, for which simulations show good properties for the higher order methods, except near the origin, for reasons discussed in relation to the next example. Similar comments apply to the Gaussian mixture.

The numerical difficulties mentioned above arise with the Student- t example. One approach to dealing with them is to adapt the null hypothesis without affecting the limiting approximations. In conventional settings, the limiting power for testing $\psi = \psi_0$ against $\psi_1 = \psi_0 + n^{-1/2}\delta$ depends on δ only, so we may argue heuristically that tests of ψ_0 or of $\psi'_0 = \psi_0 + bn^{-(1+\epsilon)/2}$ for some positive ϵ and b chosen so that ψ'_0 is interior to the parameter space should be indistinguishable in large samples. Table 5 shows coverages for right-tailed confidence intervals for testing $\psi = 1/\nu = n^{-(1+\epsilon)/2}$ in the Student- t example, for $\epsilon = 0.2, 0.4,$ and 1 . The boundary is approached more rapidly for larger ϵ , but, on the other hand, the error introduced in the asymptotic approximation will be smaller, so one might expect a trade-off between the resulting properties. Taking $\epsilon = 0.2$ and using r^* gives surprisingly good coverages, at least in the distribution tail, but larger values of ϵ lead to deterioration of the approximation for larger significance levels

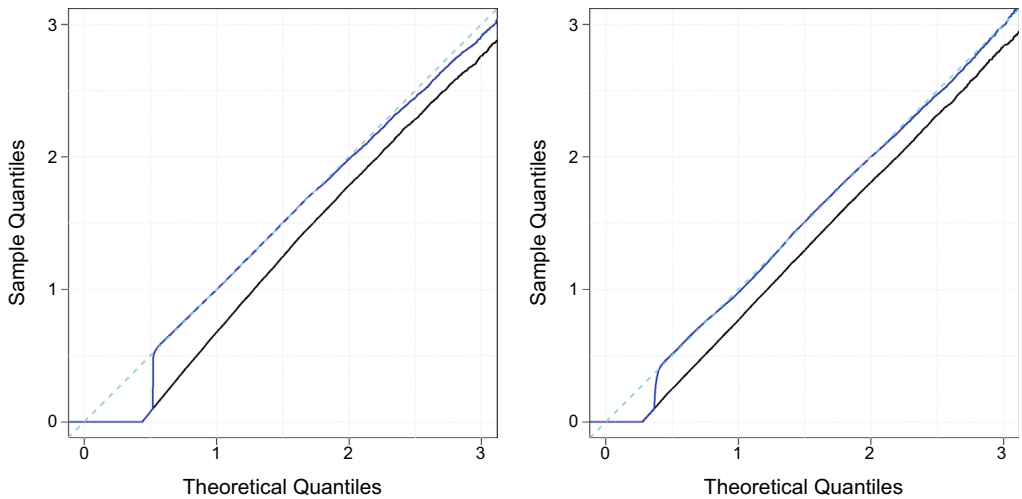


FIGURE 1: Gaussian Q–Q plots of the positive values of the likelihood root (black) and modified likelihood root (blue) based on 10^4 simulated random samples of sizes $n = 20$ (left) and $n = 100$ (right) of Student- t data with $\psi = n^{-(1+\epsilon)/2}$ reciprocal degrees of freedom for $\epsilon = 0.2$.

α . Figure 1 explains this phenomenon: the higher order approximation works very well when $\hat{\psi} > 0$, but if the probability of this event is smaller than 0.5, then the quantile–quantile plot has an awkward kink near the origin. In practical terms, this is unimportant, since we hope for reliable inferences when the test statistic lies in the upper tail of its distribution, i.e., α is small, corresponding to the upper right of the plots. Inferences based on r are conservative, because its distribution is shifted left relative to the standard normal.

Our overall conclusion from these examples is that higher order methods can produce better tail approximations in such settings, with the caveat that it may be necessary to adjust the null hypothesis to avoid numerical issues on a hard boundary and that large p -values may be unreliable. Fortunately, one is typically concerned about inference from smaller p -values in applications.

5. LINEAR MIXED MODELS

5.1. Model Setup

The variance components example of Section 2.2 can be written as a linear mixed model

$$y|b \sim \mathcal{N}_n(X\beta + Zb, \sigma^2 I_n) \quad b \sim \mathcal{N}_q(0, \sigma_b^2 \Delta), \quad (14)$$

where $\mathcal{N}_m(a, B)$ denotes the m -dimensional normal distribution with mean vector a and covariance matrix B ; X and Z are known $n \times p$ and $n \times q$ matrices of explanatory variables; β is a $p \times 1$ vector of parameters; b is a $q \times 1$ vector of “random effects” whose correlations are determined by the known $q \times q$ symmetric positive definite matrix Δ ; and σ^2 and σ_b^2 are nonnegative variance parameters. This formulation encompasses models such as the one identified in Equation (3), in which the columns of Z consist of indicators for membership of the different groups and $\Delta = I_q$, but also contains models in which smooth curves are fitted using basis functions such as smoothing splines; in this case, the columns of Z correspond to basis functions, and smoothness constraints are imposed via Δ .

If we write $\psi = \sigma_b^2/\sigma^2$ and $\Omega^{-1}(\psi) = I_n + \psi Z\Delta Z^\top$, then apart from additive constants, the log-likelihood function $\ell(\psi, \beta, \sigma^2)$ based on the marginal $\mathcal{N}_n\{X\beta, \sigma^2\Omega^{-1}(\psi)\}$ distribution of y is

$$-\frac{1}{2} \left\{ n \log \sigma^2 - \log |\Omega(\psi)| + \frac{1}{\sigma^2} (y - X\beta)^\top \Omega(\psi) (y - X\beta) \right\} \quad (15)$$

for $\psi \geq 0$, so the maximum likelihood estimates of σ^2 and β for fixed ψ are

$$\hat{\sigma}_\psi^2 = \frac{1}{n} (y - X\hat{\beta}_\psi)^\top \Omega(\psi) (y - X\hat{\beta}_\psi), \quad (16)$$

$$\hat{\beta}_\psi = \{X^\top \Omega(\psi) X\}^{-1} X^\top \Omega(\psi) y, \quad (17)$$

and the profile log-likelihood function for ψ is

$$\ell_p(\psi) \equiv -\frac{1}{2} \left\{ n \log \hat{\sigma}_\psi^2 - \log |\Omega(\psi)| \right\} \quad \text{for } \psi \geq 0. \quad (18)$$

Likewise, the log restricted likelihood (Harville, 1977) is

$$\ell_R(\psi, \sigma^2) \equiv \ell(\psi, \hat{\beta}_\psi, \sigma^2) + \frac{1}{2} \left\{ p \log \sigma^2 - \log |X^\top \Omega(\psi) X| \right\}, \quad (19)$$

where $\hat{\beta}_\psi$ is defined in Equation (17), and the corresponding estimator of σ^2 is $\hat{\sigma}_{\psi,R}^2 = (y - X\hat{\beta}_\psi)^\top \Omega(\psi) (y - X\hat{\beta}_\psi) / (n - p)$.

The crude approach to inference on ψ based on Equation (8) and under the assumption that $\psi = 0$ requires

$$\begin{aligned} p_+ &= pr_0(\hat{\psi} > 0) = pr_0 \left\{ \left. \frac{\partial \ell_p(\psi)}{\partial \psi} \right|_{\psi=0} > 0 \right\} \\ &= pr_0 \left\{ \frac{y^\top (I - H) Z Z^\top (I - H) y}{y^\top (I - H) y} > \frac{\text{tr}(Z Z^\top)}{n} \right\} \\ &= pr_0(\varepsilon^\top Q \varepsilon > 0), \end{aligned} \quad (20)$$

where $H = X(X^\top X)^{-1} X^\top$, $\text{tr}(\cdot)$ denotes the trace of a matrix, $\varepsilon \sim \mathcal{N}_n(0, I_n)$, and

$$Q = (I - H) \{n Z Z^\top - \text{tr}(Z^\top Z) I_n\} (I - H).$$

Hence, $p_+ = pr\left(\sum_{j=1}^n q_j \varepsilon_j^2 > 0\right)$, where $q_1 \geq \dots \geq q_n$ are the eigenvalues of Q . Kuonen (1999) proposed a saddlepoint approximation to this probability that is as fast as exact methods, almost as accurate, and much easier to program. The corresponding probability based on the restricted likelihood can be written in the same form, with a different matrix Q . Numerical experiments show that the saddlepoint approximation for the cases in Table 1 differs by at most 0.002 from the exact probabilities and that it improves on the χ^2 approximations of Pearson (1959) and Imhof (1961), though the latter is also quite accurate. Hence, using Equation (7) is reasonable when the likelihood roots based on Equation (15) or, better, Equation (19), are approximately normal.

Although such approximations may be worthwhile in some cases, the boundary $\psi = 0$ is soft, as the log-likelihood specified in Equation (15) is mathematically well defined provided $\Omega^{-1}(\psi)$ is positive definite, which is the case provided $\psi > -1/d_n$, where $d_n > 0$ is the largest eigenvalue of $Z\Delta Z^T$. Similar comments apply to the restricted log-likelihood function identified in Equation (19). Hence, when $\psi = 0$ lies in the interior of the “mathematical” parameter space, no difficulties with the existence of derivatives arise on the statistical boundary, and higher order methods can be applied. Skovgaard (1996) proposed a higher order approximation that differs from the tangent exponential model and verified its high accuracy when applied to the mean in a one-way random effects model. Lyons & Peters (2000) investigated this approximation more broadly in linear mixed-effects models and found it highly accurate, including for tests of zero variance parameters. When revising this article, we became aware of Iglesias-Gonzalez (2007), a University of Toronto PhD thesis by Sigfrido Iglesias-Gonzalez completed under the supervision of Nancy Reid, which investigated higher order methods for the linear mixed model in detail, including parallel derivations of elements for the tangent exponential model. None of these earlier works considered applications to smooth regression models.

The Appendix outlines the elements needed to apply the tangent exponential approximation described in Section 4.1 to the linear mixed model.

5.2. Penalized Spline Models

We now construct highly accurate tests of the hypothesis that it is not necessary to include semiparametric terms in a regression model. A common situation is when the model is of the form identified by Equation (14), the matrix X has columns corresponding to linear regression on an explanatory variable x , and the columns of Z correspond to basis functions such as natural cubic splines, B-splines, or similar. Testing for adequacy of the linear model therefore amounts to testing the soft boundary hypothesis that $\psi = \sigma_b^2/\sigma^2$ equals zero.

Such models are widely used, with the choice of basis functions and penalty usually determined by the complexity of the problem at hand and of the underlying data (de Boor, 1978; Hastie & Tibshirani, 1990; Wahba, 1990; Green & Silverman, 1994; Wood, 2017). For example, Ruppert, Wand & Carroll (2003) discuss the use of truncated power bases of order p , in which the mean for a response at a scalar x equals

$$\mu(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^K b_k (x - \kappa_k)_+^p, \quad (21)$$

where $a_+ = \max(a, 0)$, and $\kappa_1, \dots, \kappa_K$ are the knots of the spline. Some care is needed in the choice of the covariance matrix for b_1, \dots, b_K however, as a symmetric matrix will imply that the variance of the spline part increases with x , which may not be plausible. Crainiceanu, Ruppert & Vogelsang (2002) investigated testing for this variance to equal zero when $p = 0$, corresponding to a constant mean, and $p = 1$, corresponding to a linear polynomial, and found that p_+ was smaller than 0.1 for any choice of knots. We prefer natural cubic splines, which lead to mean functions $\mu(x)$ whose variance is roughly constant. Table 6 shows values of p_+ obtained using the Imhof and saddlepoint approximations, and the likelihood root and modified likelihood root, for a setting with n equally spaced knots in the unit interval. The values of p_+ are noticeably smaller than 0.5 for all of the approximations except the modified likelihood root r^* , and quantile–quantile plots confirm that the distribution of r^* is remarkably close to the standard normal.

5.3. Illustration

Figure 2 shows the 10 highest sea levels recorded in the city of Venice over the period 1887–2017; for 1935 only the six highest values are available. To illustrate the approximations described

TABLE 6: Probability p_+ (%) of positive estimates using Imhof and saddlepoint approximations, and 10^4 samples of the pivots r and r^* .

n	Imhof (Usual)	Saddlepoint (Usual)	Imhof (REML)	Saddlepoint (REML)	r (Usual)	r^* (Usual)
20	36.5	33.6	36.9	33.3	38.2	50.6
50	33.8	32.7	34.0	32.3	34.9	50.6
100	33.0	32.4	33.1	32.0	34.7	50.7
200	32.6	32.3	32.7	31.8	33.7	49.9

Note: Bold figures are within simulation error of 50%.

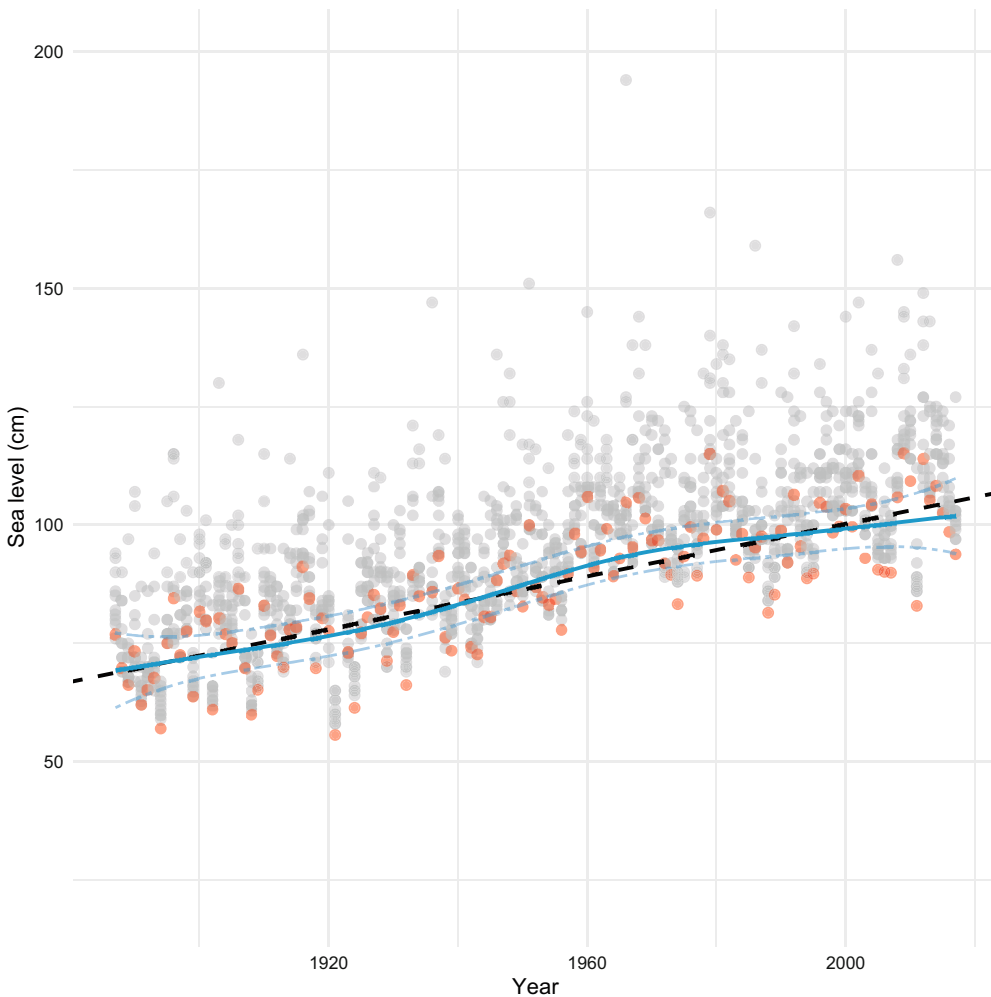


FIGURE 2: Annual average (pink) of 10 highest sea levels (cm) in Venice annually from 1887 to 2017 (grey) (only the 6 highest values are available in 1935), with linear fit (black dashed line) and cubic spline fit (light blue) and its 95% pointwise confidence interval (dot-dashed light blue).

above, we check whether a simple linear model is an adequate description of the annual averages of these observations. After a natural cubic spline fit with the degrees of freedom chosen by maximum likelihood, the values of r and r^* , respectively, 1.067 and 1.471 lead to P -values of 0.14 and 0.07 based on the limiting standard normal distributions of these statistics. As expected, the inference based on r^* is more stringent, though it does not change the conclusion in this case.

6. CONCLUSIONS

Large-sample approximations for testing a boundary hypothesis can be very poor because the weights appearing in the limiting distribution given in Equation (2) can be wildly inappropriate in practice. Even in quite large samples, the probability that the maximum likelihood estimator of a single boundary parameter equals zero can be much larger than the limiting probability. The corresponding “iceberg estimator” seems to have a downward bias in numerous examples, and it would be helpful to understand why.

Approaches to improved inference in this simplest possible case are not equally useful. Direct correction of the probability of a positive maximum likelihood estimator of the boundary parameter can be based on the profile score at both hard and soft boundaries, but the analytical computations are unpleasant and such corrections are only moderately useful.

Higher order approximation appears to work essentially perfectly for soft boundaries in the cases we have considered, and our experiments with moving the null hypothesis to lie at an asymptotically negligible distance away from a hard boundary suggest that this remedy can also be applied in other settings. More work is needed to understand how this trick affects the power of the test and to suggest appropriate choices of its parameters.

In more complex settings, bootstrap simulation under the null hypothesis seems attractive, and this would often be sensible in one-off analyses in which the programming and computational overhead are manageable. For routine use, however, it would be valuable to have analytical results to supplement standard computer output for fits of splines in additive models and their generalizations.

It would be worthwhile to investigate how higher order methods could be applied in situations with several boundary parameters.

ACKNOWLEDGEMENTS

We thank the reviewers for very helpful comments. This work was funded by the Swiss National Science Foundation and by the 2018–2022 Project of Excellence awarded to the Department of Statistical Sciences, University of Padova, by the Italian Ministry of Education, Universities and Research. Open access funding provided by Ecole Polytechnique Fédérale de Lausanne.

REFERENCES

- Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, 73, 307–322.
- Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics*, Chapman & Hall, London.
- Brazzale, A. R., Davison, A. C., & Reid, N. (2007). *Applied Asymptotics: Case Studies in Small Sample Statistics*, Cambridge University Press, Cambridge.
- Brazzale, A. R. & Mameli, V. (2023). Likelihood asymptotics in nonregular settings: A review with emphasis on the likelihood ratio. arXiv preprint, arXiv:2206.15178.
- Castillo, J. D. & López-Ratera, A. (2006). Saddlepoint approximation in exponential models with boundary points. *Bernoulli*, 12, 491–500.
- Cheng, R. C. H. & Traylor, L. (1995). Non-regular maximum likelihood problems (with discussion). *Journal of the Royal Statistical Society: Series B*, 57, 3–44.
- Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25, 573–578.

- Chernoff, H. (1956). Large-sample theory: Parametric case. *Annals of Mathematical Statistics*, 27, 1–22.
- Cox, D. R. (2006). *Principles of Statistical Inference*, Cambridge University Press, Cambridge.
- Crainiceanu, C. M., Ruppert, D. & Vogelsang, T. J. (2002). Probability that the MLE of a variance component is zero with applications to likelihood ratio tests. <http://www.orie.cornell.edu/~davidr/papers>
- Cramér, H. (1946). *Mathematical Methods of Statistics*, Princeton University Press, Princeton.
- Davison, A. C. (2003). *Statistical Models*, Cambridge University Press, Cambridge.
- Davison, A. C. & Reid, N. (2023). The tangent exponential model. In Berger, J. O., Meng, X.-L., Reid, N., & Xie, M. (Eds.), *Handbook of Bayesian, Fiducial and Frequentist Inference*, Chapman & Hall/CRC, Boca Raton.
- de Boor, C. (1978). *A Practical Guide to Splines*, Springer, New York.
- Dietz, E. & Böhning, D. (1995). Statistical inference based on a general model of unobserved heterogeneity. In Seeber, G. U. H., Francis, B. J., Hatzinger, R., & Steckel-Berger, G. (Eds.), *Statistical Modelling*, Springer, New York, 75–82.
- Elkantassi, S. (2023). *Higher Order Asymptotics: Applications to Satellite Conjunction and Boundary Problems*. Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.
- Feng, Z. & McCulloch, C. E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics & Probability Letters*, 13, 325–332.
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London: Series A*, 222, 309–368.
- Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.
- Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London: Series A*, 144, 285–307.
- Fraser, D. A. S. (2017). *p*-values: The insight to modern statistical inference. *Annual Review of Statistics and its Application*, 4, 1–14.
- Fraser, D. A. S. & Reid, N. (1993). Third order asymptotic models: Likelihood functions leading to accurate approximations to distribution functions. *Statistica Sinica*, 3, 67–82.
- Fraser, D. A. S. & Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematica*, 47, 33–53.
- Fraser, D. A. S., Reid, N., & Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika*, 86, 249–264.
- Geyer, C. J. (1994). On the asymptotics of constrained *M*-estimation. *Annals of Statistics*, 22, 1993–2010.
- Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, London.
- Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320–338.
- Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman & Hall, London.
- Iglesias-Gonzalez, S. (2007). *Highly Accurate Tests for the Mixed Linear Model*. Ph.D. thesis, University of Toronto, Toronto, Canada.
- Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48, 352–363.
- Kopylev, L. (2012). Constrained parameters in applications: Review of issues and approaches. *International Scholarly Research Network: Biomathematics*, 2012, 872956.
- Kopylev, L. & Sinha, B. (2011). On the asymptotic distribution of likelihood ratio test when parameters lie on the boundary. *Sankhyā B*, 73, 20–41.
- Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, 50, 403–418.
- Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, 86, 929–935.
- Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.
- Lyons, B. & Peters, D. (2000). Applying Skovgaard's modified directed likelihood statistic to mixed linear models. *Journal of Statistical Computation and Simulation*, 65, 225–242.
- McCullagh, P. & Tibshirani, R. J. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society: Series B*, 52, 325–344.

- McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*, Wiley, New York.
- Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford.
- Pearson, E. S. (1959). Note on an approximation to the distribution of non-central χ^2 . *Biometrika*, 46, 364.
- Reid, N. (2003). Asymptotics and the theory of inference. *Annals of Statistics*, 31, 1695–1731.
- Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.
- Self, S. G. & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82, 605–610.
- Severini, T. A. (2000). *Likelihood Methods in Statistics*, Clarendon Press, Oxford.
- Sinha, B. K., Kopylev, L., & Fox, J. (2012). Some new aspects of statistical inference for multistage dose-response models with applications. *Pakistan Journal of Statistics and Operation Research*, 8, 441–478.
- Skovgaard, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli*, 2, 145–166.
- Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72, 67–92.
- Smith, R. L. (1989). A survey of nonregular problems. *Bulletin of the International Statistical Institute*, 53, 353–372.
- Wahba, G. (1990). *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia.
- Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd ed., Chapman and Hall/CRC, Boca Raton.

APPENDIX

This appendix contains the detailed computations for the tangent exponential approximation to the variance components and linear mixed-effects models.

Variance Components

For the one-way classification model and the restricted likelihood function, the parameter vector and data are $\theta = (\psi, \sigma^2)$ and $y = (y_1, y_2) = (C_1, C_2)$ in the notation of Section 2.2. Using the pivots

$$z_1(\theta, y) = \frac{C_1}{\sigma^2(1 + m\psi)}, \quad z_2(\theta, y) = \frac{C_2}{\sigma^2}$$

to define the sufficient directions gives

$$v_\psi = \left\{ \frac{m}{m-1} \frac{k-1}{k} C_2, k(m-1) \frac{C_1}{C_2} \right\}, \quad v_{\sigma^2} = \{0, k(m-1)\}.$$

The likelihood derivatives with respect to the data are

$$\frac{\partial \ell_R}{\partial y_1} = -\frac{1}{2\sigma^2(1 + m\psi)}, \quad \frac{\partial \ell_R}{\partial y_2} = -\frac{1}{2\sigma^2},$$

which are weighted by the columns of V in order to produce the canonical parameter $\varphi(\theta) = (\varphi_1(\theta), \varphi_2(\theta))^\top$, where

$$\varphi_1(\theta) = -\frac{1}{2\sigma^2} \frac{m}{m-1} \frac{k-1}{k} \frac{C_2}{1 + m\psi},$$

$$\varphi_2(\theta) = -\frac{1}{2\sigma^2} k(m-1) \left(1 + \frac{1}{1 + m\psi} \frac{C_1}{C_2} \right).$$

The observed information matrix evaluated at the restricted maximum likelihood (REML) estimator is

$$j(\hat{\theta}) = \frac{1}{2} \begin{bmatrix} \frac{m^2}{(m-1)^2} \frac{(k-1)^3}{k^2} \left(\frac{C_2}{C_1} \right)^2 & \frac{m(k-1)^2}{C_1} \\ \frac{m(k-1)^2}{C_1} & \frac{k^2(m-1)^2(km-1)}{C_2^2} \end{bmatrix}.$$

The other quantity needed to compute $q(\psi)$ is the matrix

$$\varphi_{\theta}(\theta) = \frac{1}{2\sigma^2} \begin{bmatrix} \frac{m^2}{m-1} \frac{k-1}{k} \frac{C_2}{(1+m\psi)^2} & \frac{1}{\sigma^2} \frac{m}{m-1} \frac{k-1}{k} \frac{C_2}{(1+m\psi)} \\ \frac{mk(m-1)}{(1+m\psi)^2} \frac{C_1}{C_2} & \frac{k(m-1)}{\sigma^2} \left(1 + \frac{C_1}{C_2(1+m\psi)} \right) \end{bmatrix},$$

the second column of which contains $\varphi_{\lambda}(\theta)$; this should be evaluated at $\hat{\sigma}_{\psi}^2 = (km-1)^{-1} \{C_2 + C_1/(1+m\psi)\}$. Combining the above computations yields

$$\begin{aligned} r(\psi) &= \text{sign}(\hat{\psi} - \psi) [(km-1) \log \{C_1 + C_2(1+m\psi)\} \\ &\quad - k(m-1) \log(1+m\psi)]^{1/2}, \\ q(\psi) &= \frac{k(m-1)C_1 - (k-1)(1+m\psi)C_2}{C_1 + C_2(1+m\psi)} \left\{ \frac{1}{2} \frac{(km-1)}{k(k-1)(m-1)} \right\}^{1/2}. \end{aligned}$$

If instead we consider the ordinary likelihood function, we use the further pivot $z_3(\theta, y) = km(\bar{y}_{..} - \mu)^2 / \{\sigma^2(1+m\psi)\}$. Similar computations then allow us to express $r(\psi)$ as

$$\text{sign}(\hat{\psi} - \psi) [mk \log \{C_1 + C_2(1+m\psi)\} - k(m-1) \log(1+m\psi)]^{1/2}$$

and $q(\psi)$ as

$$m \{(m-1)C_1 - C_2(1+m\psi)\} \left[\frac{kC_1}{2(m-1)\{C_1 + C_2(1+m\psi)\}^3} \right]^{1/2}.$$

Linear Mixed Models

In a more general context, as in the model defined in Equation (14), for $\theta = (\psi, \beta, \sigma^2)$, we take the pivots to be the elements of the $n \times 1$ vector of scaled martingale differences $z(y; \theta) = \Delta(\psi)^{1/2}(y - X\beta)/\sigma$, which have independent standard normal distributions. The derivatives of the pivotal quantities are

$$v_{\beta} = X, \quad v_{\sigma^2} = \frac{(y - X\beta)}{2\sigma^2} \Big|_{(y^0, \hat{\theta}^0)}.$$

To obtain the column vector associated with ψ , we consider the eigenvalue decomposition of $\Delta(\psi)^{-1} = P\Lambda(\psi)P^T$, where P is independent of ψ , and $\Lambda(\psi)$ is an $n \times n$ diagonal matrix with elements $1 + \psi\lambda_j$, leading to

$$\frac{\partial z}{\partial \psi} = P\tilde{\Lambda}(\psi)P^T, \quad \tilde{\Lambda}(\psi) = \text{diag} \left[\frac{\lambda_1}{2(1 + \psi\lambda_1)^{3/2}}, \dots, \frac{\lambda_n}{2(1 + \psi\lambda_n)^{3/2}} \right].$$

Hence

$$v_{\psi} = P\Lambda(\psi)^{1/2}\tilde{\Lambda}(\psi)P(y - X\beta) \Big|_{(y^0, \hat{\theta}^0)}.$$

The partial derivative of the ordinary log-likelihood function with respect to the data,

$$\left. \frac{\partial \ell(\theta)}{\partial y} \right|_{y=y^o} = -\frac{1}{2\sigma^2} (y^o - X\beta)^\top \Delta(\psi),$$

is the final component needed to compute the local parameterization $\varphi(\theta)$.

For the corresponding restricted likelihood function, we take the pivotal quantities to be

$$\sigma^{-1} \Sigma(\psi)^{-1/2} e \sim \mathcal{N}_n(0, I_n),$$

where e is the residual vector $e = y - X\hat{B} = H(\psi)y$, with

$$H(\psi) = I_n - X \{X^\top \Delta(\psi) X\}^{-1} X^\top \Delta(\psi),$$

and $\Sigma(\psi) = H(\psi) \Delta(\psi)^{-1} H(\psi)^\top$, and then proceed as for the ordinary likelihood function.

Received 1 May 2023

Accepted 30 May 2023