ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 5, *pp.* 1349–1369

Modified likelihood root in high dimensions

Yanbo Tang

University of Toronto and Vector Institute, Toronto, Canada

and Nancy Reid

University of Toronto, Canada

[Received February 2020. Revised June 2020]

Summary. We examine a higher order approximation to the significance function with increasing numbers of nuisance parameters, based on the normal approximation to an adjusted log-likelihood root. We show that the rate of the correction for nuisance parameters is larger than the correction for non-normality, when the parameter dimension p is $O(n^{\alpha})$ for $\alpha < \frac{1}{2}$. We specialize the results to linear exponential families and location–scale families and illustrate these with simulations.

Keywords: High dimensional data; Higher order asymptotics; Modified likelihood root; Nuisance parameter; Statistical inference

1. Introduction

There is a growing literature on the asymptotic theory for likelihood-based inference in the high dimensional setting with the number of parameters, p, scaling with the number of observations, n, where the classical asymptotic theory fails. Sur *et al.* (2019) derived the limiting distribution of the log-likelihood ratio statistic in logistic regression, and Sur and Candès (2019) showed that the limiting distribution of the maximum likelihood estimator of the regression parameter is normal, but with expected value and asymptotic variance different from the fixed p setting. Numerical work in Sur *et al.* (2019) and Sur and Candès (2019) compares the new theory with the usual first-order approximations; for example Sur *et al.* (2019) showed that their newly derived approximation to the distribution of the log-likelihood ratio statistic is more accurate than the classical χ^2-approximation.

As emphasized in Cox (1988) an important objective of asymptotic theory is to provide approximations to distributions of inferential summaries for applications. Results motivated by an asymptotic theory with $n \to \infty$ are used to provide approximations for fixed values of n. The adequacy of the approximations is usually studied in simulations, as precise bounds are rarely available. There is a long history of developing improved approximations using asymptotic expansions, rather than simply relying on the limiting distribution. Bartlett (1937) showed that the χ^2-approximation to the distribution of the log-likelihoood ratio statistic for testing the equality of several normal variances could be much improved by a simple rescaling by the leading term in the expansion of its expected value. Use of this technique in general models, developed in Lawley (1956) following Bartlett (1953), is now usually called Bartlett correction.

Address for correspondence: Yanbo Tang, Department of Statistical Sciences, University of Toronto, Sidney Smith Hall, Toronto, Ontario, M5S 1A1, Canada.
E-mail: yanbo@utstat.toronto.edu

 1369–7412/20/821349

When the parameter of interest is a scalar, a very accurate approximation to the significance function for that parameter is the normal approximation to the distribution of an adjusted version of the signed square root of the log-likelihood ratio statistic: the modified likelihood root, r^* (Barndorff-Nielsen and Cox (1994), chapter 3). Although this approximation is asymptotically equivalent to Bartlett correction of the log-likelihood ratio statistic, it preserves the direction of the departure and has been observed in simulations to give more accurate approximations. It has also been observed in simulations that this approximation adjusts quite effectively for even large numbers of nuisance parameters. This suggests that for any fixed values of p and n it would be of interest to compare the approximations developed in the increasing p setting with higher order approximations for p fixed.

As a step in that direction, we consider the asymptotic behaviour of r^* with p increasing with n, under the constraint that $p = O(n^\alpha), \alpha < \frac{1}{2}$. This constraint on p was assumed in Fan *et al.* (2019) for generalized linear models. Sur *et al.* (2019) and Sur and Candès (2019) studied likelihood inference under the weaker condition, $p/n \to \kappa$ for some bounded constant; they referred to this as the moderate dimension setting. Of course in any given (n, p) setting, we would not usually know which of these regimes applies. One motivation for the current work is to try to assess the effect of the number of nuisance parameters on higher order approximations, to gain some insight into when the newly developed asymptotic theory is necessary for accurate approximation, and when inference based on the normal approximation to the distribution of r^* would be adequate.

Fan *et al.* (2019) showed for logistic regression, under some technical assumptions, that it is possible for the standard first-order techniques to perform adequately for $p = o(n^{1/2})$, but not for faster scaling of p. Sartori (2003) studied likelihood-based inference in Neyman–Scott models, in which observations are collected in strata, with a common parameter of interest across strata, and separate nuisance parameters in each stratum. He showed that inference based on profile likelihood was valid for $p = o(n^{1/2})$, whereas that based on modified profile likelihood was valid for relatively larger $p = o(n^{2/3})$. Portnoy (1988) developed asymptotic theory for increasing p in linear exponential families. Shun and McCullagh (1995) developed Laplace approximations in the high dimensional setting by using formal expansions and showed that for a regression model in the exponential family p must scale at a rate of $n^{1/3}$ for the approximation error to be $o(1)$, under some assumptions on the cumulants of the observations.

We examine the behaviour of the normal approximation to the distribution of r^* by considering separately two components of the modification: the nuisance parameter adjustment r_{np}, and the information adjustment r_{inf}, as in Pierce and Peters (1992). We show that under smoothness assumptions on the model, for $p = O(n^\alpha)$ for $\alpha < \frac{1}{2}$, that $r_{\mathrm{inf}} = O_p(p^{3/2}/n^{1/2})$ and $r_{\mathrm{np}} = O_p(p/n^{1/2})$. Thus r^* behaves as in the classical asymptotic regime for $p = o(n^{1/3})$. As noted by a reviewer, this rate is suggested by the fixed p expansions of the maximum likelihood estimator in Barndorff-Nielsen and Cox (1994), section 5.3. The difference in the scaling rates of r_{inf} and r_{np} explains long-standing empirical evidence that the nuisance parameter adjustment plays a larger role than the information adjustment. As a by-product of some of the intermediate expansions that are used in the proof of the main results, we quantify the deviation of a general model from a linear exponential model. This is useful as the expression for r^* in the linear exponential family has a particularly simple form.

Specializing the results to linear exponential models or location–scale models we show that $r_{\mathrm{inf}} = O_p(n^{-1/2})$ and $r_{\mathrm{np}} = O_p(pn^{-1/2})$, which implies that, for $p = O(n^{1/2})$, the nuisance adjustment is asymptotically non-negligible. Simulations suggest that the normal approximation to the distribution of the likelihood root r breaks down at $p = O(n^{1/2})$, whereas the normal approximation to the distribution of r^* breaks down at $p = O(n^{2/3})$. We also briefly discuss the Bayesian version of r^* that is obtained via the Laplace expansion.

We establish a general upper bound on the rate of growth of r_{np} and r_{inf} but do not establish any lower bounds, so our results may be pessimistic for some specific models, as is suggested in Section 6. An ideal analysis would give the exact divergence point between the first-order and the higher order approximations, by establishing sharp rates of growth of r_{inf} and r_{np} in p and n, e.g. by establishing matching upper and lower bounds, but this seems quite difficult. The numerical work in Section 6 suggests that these bounds would depend on both the model and the parameter values.

We would also like to be able to confirm that the normal approximation to the distribution of r^* is more accurate than the normal approximation to the distribution of r, even as p increases with n. This would require extension of the p^*-approximation to the high dimensional setting. A brief discussion of this point is given in Section 7.

In what follows we use the following notation. For a vector u, we let $\|u\|_2$ denote the Euclidean norm of u. With A an $m \times n$ matrix with (i, j) entry a_{ij}, the ordered singular values of A are $\eta_1(A) \geqslant \eta_2(A) \geqslant \ldots \geqslant \eta_{\min(n,m)}(A)$. The operator norm is

$$\|A\|_{\mathrm{op}} = \eta_1(A),$$

and the Frobenius norm is

$$\|A\|_{\mathrm{F}} = \mathrm{tr}(A^{\mathrm{T}}A)^{1/2} = (\sum_{i,j} a_{ij}^2)^{1/2}.$$

In general

$$\|A\|_{\mathrm{op}} \leqslant \|A\|_{\mathrm{F}}, \tag{1}$$

and the Frobenius norm and the operator norm are equal if A is a vector. For two matrices A and B of compatible dimension

$$|\mathrm{tr}(A^{\mathrm{T}}B)| \leqslant \{\mathrm{tr}(A^{\mathrm{T}}A)\mathrm{tr}(B^{\mathrm{T}}B)\}^{1/2} = \|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}.$$

For $p \times p$ matrices A and B, von Neumann's trace inequality is

$$|\mathrm{tr}(AB)| \leqslant \sum_{j=1}^{p} \eta_j(A)\eta_j(B).$$

We sometimes make assumptions about the largest singular value of a matrix A, but otherwise we use inequality (1) to provide an upper bound on the largest singular value.

The programs that were used to carry out the simulations can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets`.

2. Higher order approximations: definitions and background

We assume a parametric model for independent observations $y = (y_1, \ldots, y_n)^{\mathrm{T}}$, where the distribution of y_i has density $f(\cdot; \theta)$ depending on a set of unknown parameters θ of dimension p, and possibly on a vector of covariates x_i. We consider inference for a one-dimensional parameter of interest, ψ, and write $\theta = (\psi, \lambda)$, where λ is a $(p-1)$-dimensional vector of nuisance parameters.

The observed Fisher information function is $j(\theta) = -\partial^2 l(\theta)/\partial\theta\partial\theta^{\mathrm{T}}$, and subscripts on $j(\theta)$ denote subblocks of this matrix. Other derivatives of the log-likelihood function are denoted by subscripts on $l(\theta)$; for example $l_{\psi\lambda\lambda}(\theta)$ is the matrix with entries $\{l_{\psi\lambda\lambda}(\theta)\}_{rs} = \partial^3 l(\theta)/\partial\psi\partial\lambda_r\partial\lambda_s$.

The profile log-likelihood function is

$$l_{\mathrm{p}}(\psi) = \sup_{\lambda} l(\psi, \lambda) = l(\psi, \hat{\lambda}_{\psi}),$$

where $\hat{\lambda}_{\psi}$ is the constrained maximum likelihood estimator, and the profile information function is $j_{\mathrm{p}}(\psi) = -l''_{\mathrm{p}}(\psi)$. When p is fixed, the profile log-likelihood ratio statistic

$$w(\psi_0) = 2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)\}$$

converges in distribution under the model $f(y; \theta_0)$ to a χ^2_1 random variable, under regularity conditions on the model (Barndorff-Nielsen and Cox (1994), chapter 2). The directed root

$$r(\psi_0) = \mathrm{sgn}(\hat{\psi} - \psi_0)[2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)\}]^{1/2}$$

converges in distribution under the same conditions to an $N(0, 1)$ random variable. The standard normal approximation to the distribution of $r(\psi_0)$ has relative error $O(n^{-1/2})$. Using higher order asymptotic expansions it can be shown that a modified version of r is more accurately approximated by the standard normal distribution. This modified directed root is

$$r^*(\psi_0) = r(\psi_0) + r^{-1}(\psi_0) \log\{u(\psi_0)/r(\psi_0)\}, \tag{2}$$

where $u(\psi_0)$ is to be defined later in equation (4). The first term in equation (2) is $O_p(1)$ and the second term is $O_p(n^{-1/2})$, and the standard normal approximation to the distribution of $r^*(\psi_0)$ has relative error $O(n^{-3/2})$ in continuous models (Barndorff-Nielsen and Cox (1994), chapter 6).

There is a discontinuity in approximation (2) at $\hat{\psi} = \psi_0$, where both r and u approach 0. In practice this means that the approximation is numerically unstable in a region near the 50% point of the distribution, although the p-value in this region is not usually of interest. In Brazzale *et al.* (2007) this region was interpolated by using a smoothing spline. In the simulations in Section 6 we omit from the summaries values of r^* when $|r| < 0.025$. We also recommend plotting the profile log-likelihood as a function of ψ to check that it is concave and unimodal in a region around $\hat{\psi}$; otherwise reliance on the asymptotic theory will be suspect.

To define the adjustment term u in equation (2), additional notation is required. We follow Barndorff-Nielsen and Cox (1994), chapter 5, and assume that the log-likelihood function $l(\theta; y) = l(\theta; \hat{\theta}, a)$ depends on the data through the maximum likelihood estimator $\hat{\theta}$ and a complementary statistic a which is either exactly or approximately ancillary. Derivatives with respect to $\hat{\theta}$ are called sample space derivatives and are denoted by $\hat{\theta}$ after a semicolon in the subscript of l: for example,

$$l_{;\hat{\theta}}(\theta) = \frac{\partial}{\partial \hat{\theta}} l(\theta; \hat{\theta}, a).$$

Derivatives of l with respect to both θ and $\hat{\theta}$ are required in the expansions below; these are referred to as mixed derivatives. Derivatives with respect to θ are denoted by placing θ before the semicolon in the subscript: for example,

$$l_{\theta;\hat{\theta}}(\theta) = \frac{\partial^2}{\partial \theta \partial \hat{\theta}} l(\theta; \hat{\theta}, a).$$

Assuming that $\hat{\theta}$ is the solution of $l_{\theta}(\hat{\theta}; \hat{\theta}, a) = 0$, differentiation with respect to $\hat{\theta}$ establishes the observed balance relation (Barndorff-Nielsen and Cox (1994), section 5.2)

$$l_{\theta;\hat{\theta}}(\hat{\theta}) = j(\hat{\theta}). \tag{3}$$

The correction term

$$u(\psi_0) = C\tilde{u} \tag{4}$$

is the product of a score-type statistic based on the profile log-likelihood function and an adjustment for nuisance parameters. The score-type statistic is

$$\tilde{u} = \tilde{u}(\psi_0) = j_{\mathrm{p}}^{-1/2}(\hat{\psi})\bar{l}_{\mathrm{p}/\hat{\psi}}(\psi_0), \tag{5}$$

where $\bar{l}_{\mathrm{p}}(\psi_0) = l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)$ and

$$\bar{l}_{\mathrm{p}/\hat{\psi}} = \{l_{;\hat{\psi}}(\hat{\theta}) - l_{;\hat{\psi}}(\hat{\theta}_{\psi_0})\} - l_{\lambda;\hat{\psi}}(\hat{\theta}_{\psi_0})l_{\lambda;\hat{\lambda}}(\hat{\theta}_{\psi_0})^{-1}\{l_{;\hat{\lambda}}(\hat{\theta}) - l_{;\hat{\lambda}}(\hat{\theta}_{\psi_0})\};$$

following Barndorff-Nielsen and Cox (1994), section 6.6, we write $\bar{l}_{\mathrm{p}/\hat{\psi}}$ to denote the sample space derivative of $\bar{l}_{\mathrm{p}}$ with respect to $\hat{\psi}$ when it is considered as a function of ψ, $\hat{\psi}$, $\hat{\lambda}_\psi$ and an approximate or exact ancillary statistic a. The adjustment for nuisance parameters is

$$C = C(\psi_0) = |l_{\lambda;\hat{\lambda}}(\hat{\theta}_{\psi_0})|/\{|j_{\lambda\lambda}(\hat{\theta}_{\psi_0})||j_{\lambda\lambda}(\hat{\theta})|\}^{1/2}. \tag{6}$$

Both C and $\tilde{u}$ are invariant under so-called interest respecting reparameterizations from (ψ, λ) to (ψ, η) where η may depend on both ψ and λ.

The modified likelihood root (2) can be decomposed accordingly, as

$$r^* = r + r_{\mathrm{np}} + r_{\mathrm{inf}},$$

where

$$\begin{aligned} r_{\mathrm{np}} &= r^{-1}\log(C), \\ r_{\mathrm{inf}} &= r^{-1}\log(\tilde{u}/r). \end{aligned} \tag{7}$$

This was suggested in Pierce and Peters (1992) in the context of linear exponential families and generalized in Barndorff-Nielsen and Cox (1994), section 6.6. Each term in expression (7) depends on the parameter of interest, ψ, and the data $(\hat{\theta}, a)$, through the log-likelihood function.

The normal approximation to the distribution of r^* as a function of $\hat{\theta}$, given a, is derived by integrating an approximation to the density of the maximum likelihood estimator (Barndorff-Nielsen, 1983). A change of variable in this integration implicitly assumes that r is a one-to-one function of $\hat{\psi}$ for fixed ψ, $\hat{\lambda}_\psi$ and a in a ball of fixed radius around ψ_0. In the case of the linear exponential family this is easily verified, and the approximation can be obtained directly by a ratio of saddlepoint approximations (Davison, 1988). More detailed discussion of the r^*-approximation and its various formulations is given in Reid (2003).

The arguments in the next section use expansions r, r_{np} and r_{inf} viewed as functions of the parameter of interest ψ. We use the mean value theorem to control the approximation error when $p = p_n$; each remainder term is evaluated at an intermediate value $\tilde{\psi}$ where $|\psi_0 - \tilde{\psi}| < |\psi_0 - \hat{\psi}|$. A similar approach is outlined in Barndorff-Nielsen and Cox (1994), section 3.3. In expansions of functions of θ, we restrict attention to a neighbourhood of θ_0 as described in Section 3.

We write $\zeta_k(\psi) = \mathrm{d}^k l_{\mathrm{p}}(\psi)/\mathrm{d}\psi^k$ and define the quasi-cumulants

$$\kappa_k(\psi) = \frac{\zeta_k(\psi)}{\{-\zeta_2(\hat{\psi})\}^{k/2}}.$$

We also define

$$\gamma_1(\psi) = \frac{\mathrm{d}}{\mathrm{d}\psi}\log\{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|\} = \mathrm{tr}\left\{ j_{\lambda\lambda}^{-1}(\psi, \hat{\lambda}_\psi)\frac{\mathrm{d}}{\mathrm{d}\psi}j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\right\}. \tag{8}$$

3. Analysis of r^* when p is increasing with n

Let $N_{\theta_0,\delta} = \{\theta : \|\theta - \theta_0\|_2 < \delta\}$ for $\delta > 0$ denote a ball of radius δ centred on θ_0. As noted above we require $p = O(n^\alpha)$ for some $0 \leqslant \alpha < \frac{1}{2}$.

The assumptions on the model are as follows.

Assumption 1. $\|\hat{\theta} - \theta_0\|_2 = o_p(1)$ and $\sup_{\psi \in A_n} \|\hat{\theta}_\psi - \theta_0\|_2 = o_p(1)$, where $A_n = \{\psi : |\psi - \psi_0| \leqslant |\hat{\psi} - \psi_0|\}$.

Assumption 2. $j_{\psi\lambda_r}(\theta) = O_p(n^{1/2})$ uniformly in r, for $\theta \in N_{\theta_0,\delta}$.

Assumption 3. The eigenvalues of $j(\theta)/n$ and $\{j(\theta)/n\}^{-1}$ are bounded in probability, for $\theta \in N_{\theta_0,\delta}$.

Assumption 4. The log-likelihood derivatives $l_{\theta_r\theta_s\theta_t}(\theta)$, $l_{\theta_r\theta_s\theta_t\theta_o}(\theta)$ and $l_{\theta_r\theta_s;\hat{\theta}_t}(\theta)$ are continuous and uniformly $O_p(n)$ in r, s, t and o, for $\theta \in N_{\theta_0,\delta}$.

Assumption 5. The log-likelihood root $r \to^{\mathrm{D}} Z$, for some random variable Z, whose distribution has no point mass at 0. The Wald statistic $t = j_{\mathrm{p}}^{1/2}(\hat{\psi})(\hat{\psi} - \psi_0) \to^{\mathrm{D}} \tilde{Z}$ for some random variable $\tilde{Z}$.

Assumption 1, norm consistency, is quite strong, as noted by a reviewer, although even in the classical setting consistency is often assumed rather than proved. Assumptions 2–4 control the behaviour of the log-likelihood derivatives and mixed derivatives when these are evaluated at $\hat{\theta}_{\tilde{\psi}}$ or $\hat{\theta}$. Assumption 2 can be satisfied by assuming that the parameter of interest is globally orthogonal or locally orthogonal in $N_{\theta_0,\delta}$ (Cox and Reid, 1987), and requiring the behaviour to be uniform for each component of λ. Assumption 3 ensures that the asymptotic covariance matrix is well behaved, and assumption 4 requires that the likelihood derivatives behave as in the fixed p setting. Assumption 5 does not require the limiting distribution of r to be the standard normal distribution but the requirement of no point mass at 0 is necessary as we need to divide some expressions by r. Finally as r_{np} and r_{inf} are invariant under interest respecting reparameterizations, the assumptions can be satisfied under any parameterization of the model.

Assumption 3 is similar to condition 2 of Fan *et al.* (2019) that the maximum and minimum eigenvalues of the rescaled observed information matrix are bounded away from 0 and ∞. Similarly conditions A1 and A3 in Lei *et al.* (2016) restrict the growth of the maximum eigenvalue of the Hessian of the objective function, in their analysis of M-estimates for high dimensional linear regression. Shun and McCullagh (1995), section 6, made assumptions similar to assumption 4 in the context of Laplace approximations in generalized linear models.

The two results below characterize the asymptotic behaviour of r_{np} and r_{inf} respectively. The proof of theorem 1 is sketched here, with technical details given in the on-line supplementary materials. The proof of theorem 2 is given in the supplementary materials.

Theorem 1. Under assumptions 1–5, $r_{\mathrm{np}} = O_p\{\max(p^{3/2}/n^{1/2}, p^3/n)\}$.

Proof. From expressions (6) and (7),

$$r_{\mathrm{np}} = \frac{1}{r} \log \left[\frac{|l_{\lambda;\hat{\lambda}}(\hat{\theta}_{\psi_0})|}{\{|j_{\lambda\lambda}(\hat{\theta}_{\psi_0})||j_{\lambda\lambda}(\hat{\theta})|\}^{1/2}} \right].$$

We have

$$|l_{\lambda;\hat{\lambda}}(\hat{\theta}_{\psi_0})| = \left| l_{\lambda;\hat{\lambda}}(\hat{\theta}) - (\hat{\psi} - \psi_0) \frac{\mathrm{d}}{\mathrm{d}\psi} l_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)|_{\hat{\theta}_{\tilde{\psi}}} \right|,$$

$$
\begin{aligned}
&= |l_{\lambda;\hat{\lambda}}(\hat{\theta}) + R_1|,\\
&= |l_{\lambda;\hat{\lambda}}(\hat{\theta})||I + l^{-1}_{\lambda;\hat{\lambda}}(\hat{\theta})R_1|,\\
&= |j_{\lambda\lambda}(\hat{\theta})||I + j^{-1}_{\lambda\lambda}(\hat{\theta})R_1|,
\end{aligned}
$$

where the final equality uses equation (3). Then

$$
r_{\mathrm{np}} = \frac{1}{r}\log\left\{\frac{|j_{\lambda\lambda}(\hat{\theta})||I + j^{-1}_{\lambda\lambda}(\hat{\theta})R_1|}{|j_{\lambda\lambda}(\hat{\theta}_{\psi_0})|^{1/2}|j_{\lambda\lambda}(\hat{\theta})|^{1/2}}\right\} = \frac{1}{r}\log\left\{\frac{|j_{\lambda\lambda}(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_{\psi_0})|^{1/2}}\right\} + \frac{1}{r}\log\{|I + j^{-1}_{\lambda\lambda}(\hat{\theta})R_1|\}, \quad (9)
$$

$$
=: \frac{1}{r}\log(\rho) + \frac{1}{r}\log\{|I + j^{-1}_{\lambda\lambda}(\hat{\theta})R_1|\}. \quad (10)
$$

In lemmas 4 and 6 in the on-line supplementary materials we show that

$$
r^{-1}\log\{|I + j^{-1}_{\lambda\lambda}(\hat{\theta})R_1|\} = O_p\left\{\max\left(\frac{p^{3/2}}{n^{1/2}}, \frac{p^3}{n}\right)\right\}.
$$

Taylor series expansion of $r^{-1}\log(\rho)$ gives

$$
\frac{1}{r}\log(\rho) = \frac{1}{2r}(\psi_0 - \hat{\psi})\gamma_1(\tilde{\psi}) = \frac{t}{2r}\frac{\gamma_1(\tilde{\psi})}{j_{\mathrm{p}}^{1/2}(\hat{\psi})},
$$

where $\gamma_1(\tilde{\psi})$ is defined in equation (8). To bound $\gamma_1(\tilde{\psi})$, we have

$$
\begin{aligned}
|\gamma_1(\tilde{\psi})| &= \left|\mathrm{tr}\left\{j^{-1}_{\lambda\lambda}(\hat{\theta}_{\tilde{\psi}})\frac{\mathrm{d}}{\mathrm{d}\psi}j_{\lambda\lambda}(\hat{\theta}_{\psi})|_{\hat{\theta}_{\tilde{\psi}}}\right\}\right| \leqslant \left(\mathrm{tr}\{j^{-1}_{\lambda\lambda}(\hat{\theta}_{\tilde{\psi}})^2\}\mathrm{tr}\left[\left\{\frac{\mathrm{d}}{\mathrm{d}\psi}j_{\lambda\lambda}(\hat{\theta}_{\psi})|_{\hat{\theta}_{\tilde{\psi}}}\right\}^2\right]\right)^{1/2},\\
&\leqslant \|j^{-1}_{\lambda\lambda}(\hat{\theta}_{\tilde{\psi}})\|_{\mathrm{F}}\left\|\frac{\mathrm{d}}{\mathrm{d}\psi}j_{\lambda\lambda}(\hat{\theta}_{\psi})|_{\hat{\theta}_{\tilde{\psi}}}\right\|_{\mathrm{F}} = O_p(p^{3/2}),
\end{aligned}
$$

since

$$
\|j^{-1}_{\lambda\lambda}(\hat{\theta}_{\tilde{\psi}})\|_{\mathrm{F}} = O_p(p^{1/2}/n),
$$

and

$$
\left\|\frac{\mathrm{d}}{\mathrm{d}\psi}j_{\lambda\lambda}(\hat{\theta}_{\psi})|_{\hat{\theta}_{\tilde{\psi}}}\right\|_{\mathrm{F}} = O_p(pn),
$$

by proposition 1 in the on-line supplementary materials, and $j^{-1}_{\mathrm{p}}(\hat{\psi}) = O_p(n^{-1})$ by assumptions 1 and 3. In lemma 3 of the on-line supplementary materials we show that $t/r = 1 + O_p(n^{-1/2})$, giving

$$
r_{\mathrm{np}} = O_p\left\{\max\left(\frac{p^{3/2}}{n^{1/2}}, \frac{p^3}{n}\right)\right\}.
$$

Remark 1. Assumption 1 may limit the possible scaling of p with n in practice, and as in the p-fixed case its verification is model dependent. Portnoy (1988) verified norm consistency for the canonical parameter in an exponential family model, and Portnoy (1984) proved the result for M-estimators in linear regression, for $p = O(n^{\alpha})$ for $\alpha < 1$. Using results from Fan *et al.* (2019), it can be shown that norm consistency holds for generalized linear models for $p = o(n^{1/2})$.

Remark 2. The quantity $\gamma_1(\psi_0)$ is the leading term of the bias of the profile score in linear exponential models in the p-fixed asymptotic regime (McCullagh and Tibshirani (1990), section 3). We show that $\gamma_1(\psi_0) = O(p)$ for the linear exponential family in the proof of proposition 1.

Thus for linear exponential models it is possible for the bias of the profile score to be unbounded in the high dimensional regime, leading to an asymptotically biased estimate of ψ_0. Kosmidis *et al.* (2019) developed procedures to debias the maximum likelihood estimator.

Remark 3. The quantity $\log(\rho)$ also appears in the modified profile likelihood function (Barndorff-Nielsen, 1983). The modification to the profile log-likelihood function in the p-fixed case is $O(1)$, but when p increases with n the modification can be quite large, as we have $\log(\rho) = O_p(p^{3/2}/n^{1/2})$. This may explain why inference based on quantities that are derived from the modified profile likelihood function is more accurate in simulations than that based on quantities that are derived from the profile likelihood function, as discussed in Sartori (2003).

Theorem 2. Under assumptions 1–5, $r_{\text{inf}} = O_p(p/n^{1/2})$.

The proof in the on-line supplementary materials uses lemmas 3 and 7.

Remark 4. The scaling rate of r_{inf} is slower than r_{np} by a factor of $p^{1/2}$, showing that the information correction has less effect on the approximation as it is asymptotically negligible for $p = o(n^{1/2})$, whereas r_{np} is only negligible for $p = o(n^{1/3})$. In the specific models that we examine in Section 5 we have the stronger result that $r_{\text{inf}} = O_p(n^{-1/2})$ so the rate does not depend on the scaling of p with n.

4. Deviation from exponentiality when p is fixed

The form of u simplifies in the linear exponential family and we have

$$r_{\text{np}} = \frac{1}{r}\log(\rho),$$
$$r_{\text{inf}} = \frac{1}{r}\log\left(\frac{t}{r}\right),$$

where t is the Wald statistic for testing $\psi = \psi_0$ and ρ is an information determinant defined in expression (10):

$$t = (\hat{\psi} - \psi_0) j_{\text{p}}^{1/2}(\hat{\psi}), \tag{11}$$

$$\rho^2 = |j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|/|j_{\lambda\lambda}(\psi_0, \hat{\lambda}_{\psi_0})|. \tag{12}$$

The expressions for r_{np} and r_{inf} in linear exponential families are easier to work with from a practical standpoint, as most statistical software provides the information matrix, Wald statistic and log-likelihood function. In general families the mixed derivatives need to be obtained for each model of interest, and this can be difficult.

In the proof of theorem 1 the intermediate result (10) shows that

$$\frac{1}{r}\log\{|I + j_{\lambda\lambda}^{-1}(\hat{\theta})R_1|\}$$

measures to $O_p(n^{-3/2})$ the deviation of r_{np} in a general model from that in an exponential family. As r_{np} has the usual order of $O_p(n^{-1/2})$, a necessary and sufficient condition for the models to be asymptotically equivalent is $r^{-1}\log\{|I + j_{\lambda\lambda}^{-1}(\hat{\theta})R_1|\} = o_p(n^{-1/2})$. This leads to the following corollary of theorem 1.

Corollary 1. When p is fixed, under assumptions 1 and 3–5 and the further assumption $\|\mathrm{d}l_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)/\mathrm{d}\psi|_\theta\|_2 = o_p(n)$, for $\theta \in N_{\theta_0,\delta}$,

$$r_{\mathrm{np}} - \frac{1}{r}\log(\rho) = \frac{1}{r}\log\{|I + j_{\lambda\lambda}^{-1}(\hat{\theta})R_1|\} = o_p(n^{-1/2}).$$

Corollary 1 implies that, if we know that the size of the third-order mixed derivative with respect to the parameter of interest is small in terms of singular value, we can use the formula for r_{np} that is associated with the linear exponential family when calculating r^*, as the difference between the two expressions is negligible. We also note that assumption 2 is automatically satisfied when p is fixed, as we may use a parameterization where the parameter of interest is orthogonal to the nuisance parameter at θ_0, and so it is not needed for this result.

Remark 5. Cox and Reid (1992) showed that the difference between the profile log-likelihood function and the adjusted version $l_{\mathrm{p}}(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\theta}_\psi)|$ is asymptotically negligible if $\mathbb{E}\{l_{\psi\lambda\lambda}(\theta_0)\} = 0$. The condition in corollary 1 on the mixed third derivative is similar. This shows as well that the condition in Cox and Reid (1992) can be weakened by using von Neumann's inequality:

$$|\mathrm{tr}\{j_{\lambda\lambda}^{-1}(\theta_0)l_{\psi\lambda\lambda}(\theta_0)\}| \leqslant \sum_{j=1}^{p} \eta_j\{j_{\lambda\lambda}^{-1}(\theta_0)\}\eta_j\{l_{\psi\lambda\lambda}(\theta_0)\} \leqslant \eta_1\{l_{\psi\lambda\lambda}(\theta_0)\}\mathrm{tr}\{j_{\lambda\lambda}^{-1}(\theta_0)\}.$$

Thus a condition on the size of the maximum singular value of $l_{\psi\lambda\lambda}(\theta_0)$ could be used in place of the condition in Cox and Reid (1992).

For r_{inf}, we have the following corollary to theorem 2.

Corollary 2. Under assumptions 1 and 3–5, and the further assumption that $l_{\psi\psi;\hat{\psi}}(\theta) = o_p(n)$ for $\theta \in N_{\theta_0,\delta}$,

$$r_{\mathrm{inf}} - \frac{1}{r}\log\left(\frac{t}{r}\right) = o_p(n^{-1/2}).$$

Remark 6. When p is fixed, a sufficient condition for both r_{inf} and r_{np} to be asymptotically equivalent to their expressions in the linear exponential family is $\|\mathrm{d}l_{\theta;\hat{\theta}}(\hat{\theta}_\psi)/\mathrm{d}\psi|_\theta\|_{\mathrm{op}} = o_p(n)$, for $\theta \in N_{\theta_0,\delta}$.

5. Examples

5.1. Linear exponential family

Let X be an $n \times p$ matrix of covariates with (i, j) entry x_{ij} and ith row x_i^{T}. We assume that the density of y_i is that of a full exponential family model with canonical parameter θ for $i = 1, \ldots, n$. The log-likelihood function for an independent sample $y_1, \ldots, y_n$ is

$$l(\psi, \lambda; y) = \psi\sum_{i=1}^{n} y_i x_{i1} + \sum_{j=2}^{p}\lambda_j\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} K(x_i^{\mathrm{T}}\theta). \tag{13}$$

Without loss of generality we assume that the parameter of interest $\psi = \theta_1$.

The parameters that are orthogonal to ψ are $\tau_j = \mathbb{E}(\Sigma_{i=1}^{n} x_{ij}y_i/n)$ for $j = 2, \ldots, p$. The constrained maximum likelihood estimate of τ does not depend on ψ, as

$$\begin{pmatrix} \Sigma y_i x_{i1} \\ \Sigma y_i x_{i2} \\ \vdots \\ \Sigma y_i x_{ip} \end{pmatrix} = \begin{pmatrix} \Sigma K_\psi(x_i^{\mathrm{T}}\hat{\theta}) \\ \Sigma K_{\lambda_1}(x_i^{\mathrm{T}}\hat{\theta}) \\ \vdots \\ \Sigma K_{\lambda_{p-1}}(x_i^{\mathrm{T}}\hat{\theta}) \end{pmatrix},$$

and the same set of equations, without the first, gives the solution of the constrained maximum likelihood estimator. In the (ψ, τ) parameterization the observed Fisher information function evaluated at the constrained maximum likelihood estimator is

$$j(\psi, \hat{\tau}) = \begin{pmatrix} \tilde{j}_{\mathrm{p}}(\psi) & 0 \\ 0 & n^2 \tilde{j}_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)^{-1} \end{pmatrix}, \tag{14}$$

where $\tilde{j}$ is the observed information function in the (ψ, λ) parameterization and $\tilde{j}_{\mathrm{p}} = j_{\psi\psi} - j_{\psi\lambda} j_{\lambda\lambda}^{-1} j_{\lambda\psi}$ is the observed profile information function in the (ψ, λ) parameterization. We also note that the expressions for r_{np} and r_{inf} no longer involve any sample space derivatives (Barndorff-Nielsen and Cox (1994), example 6.19).

Because of the simpler form of r_{np} and r_{inf} we can reformulate some of the assumptions. Assumption 2 is no longer necessary as τ is globally orthogonal to ψ. Assumptions 3 and 4 can be replaced by the following assumptions.

Assumption 6. The eigenvalues of the Gram matrix satisfy $0 < a_1 n < \eta_i(X^{\mathrm{T}} X) < a_2 n < \infty$, and $\Sigma_{i=1}^n x_{ij} x_{ik} = O(n)$ for each j and k in $(1, \ldots, p)$.

Assumption 7. $\max_{i=1,\ldots,n} K''(x_i^{\mathrm{T}}\theta) = O(1)$, $\max_{i=1,\ldots,n} \{K''(x_i^{\mathrm{T}}\theta)\}^{-1} = O(1)$ and $\Sigma_i K'''(x_i^{\mathrm{T}}\theta) x_{i1}^3 = O(n)$ for $\theta \in N_{\theta_0,\delta}$.

Assumption 8. The third log-likelihood derivative $l_{\psi\psi\psi}(\theta) = O_p(n)$, for $\theta \in N_{\theta_0,\delta}$.

Assumptions 6 and 7 imply assumption 3 and assumption 8 is a relaxation of assumption 4. We make the following additional assumption on the third derivative of the log-likelihood, which states that the observed information of the nuisance parameter must not be too sensitive to changes in the parameter of interest ψ under the orthogonal parameterization.

Assumption 9. The derivative of the observed Fisher information matrix under the (ψ, τ) parameterization with respect to ψ satisfies $\| j_{\psi\tau\tau}(\theta) \|_{\mathrm{op}} = O_p(n)$, for $\theta \in N_{\theta_0,\delta}$ for some $\delta > 0$.

Proposition 1. Under assumptions 1 and 5–9 in the linear exponential model (13)

$$r_{\mathrm{np}} = O_p(pn^{-1/2}),$$

$$r_{\mathrm{inf}} = O_p(n^{-1/2}).$$

The proof is provided in the on-line supplementary materials. Proposition 1 shows that r_{inf} has the same behaviour as in the p-fixed regime and is therefore asymptotically negligible, whereas r_{np} grows with p at a slower rate than in the general case. For $p = O(n^{1/2})$ the limiting distribution of r can differ from that of r^*, as demonstrated numerically in Section 5.

Remark 7. The score statistic $l'_{\mathrm{p}}(\psi_0) j_{\mathrm{p}}^{-1/2}(\psi_0)$ can be decomposed as

$$\frac{l'_{\mathrm{p}}(\psi_0) - E\{l'_{\mathrm{p}}(\psi_0)\}}{j_{\mathrm{p}}^{1/2}(\psi_0)} + \frac{E\{l'_{\mathrm{p}}(\psi_0)\}}{j_{\mathrm{p}}^{1/2}(\psi_0)},$$

and for $p = O(n^{1/2})$ the second term is $O_p(1)$ since as discussed in remark 2 the bias of the profile score is $O(p)$, so, even if the first term converges to a standard normal distribution, the second term can produce a non-vanishing bias. This was noted in Sartori (2003) in the context of stratified models.

Remark 8. Proposition 1 can be applied to stratified models in the linear exponential family to obtain the same scaling rates of r_{inf} and r_{np} under similar assumptions. These rates agree with those obtained by Sartori *et al.* (1999) and Portnoy (1988).

5.2. Location–scale models

We consider a linear regression model based on a location–scale family:

$$y_i = x_i^{\mathrm{T}}\beta + \sigma z_i, \tag{15}$$

where the errors z_i are assumed independent and identically distributed from a known distribution with continuous density $f(z)$.

In model (15)

$$r_{\mathrm{inf}} = \frac{1}{r}\log\left(\frac{s}{r}\right),$$
$$r_{\mathrm{np}} = -\frac{1}{r}\log(\rho),$$

where

$$s = l'_{\mathrm{p}}(\psi_0)/j_{\mathrm{p}}^{1/2}(\hat{\psi})$$

is the score statistic standardized by the observed profile information at $\hat{\psi}$. We assume that the parameter of interest is β_1 and write $(\psi, \lambda) = (\beta_1, \ldots, \beta_{p-1}, \sigma)$. We make additional assumptions on the third-derivative matrix for the location–scale model.

Assumption 10. $\max_{j=1,\ldots,p} \|j_{\theta_j\lambda\lambda}(\theta)\|_{\mathrm{op}} = O_p(n)$, for $\theta \in N_{\theta_0,\delta}$.

This assumption is needed as the derivative of the constrained maximum likelihood estimate with respect to ψ is not 0 as it is in the case of the linear exponential family. In the on-line supplementary materials we prove the following proposition.

Proposition 2. Under assumptions 1–5 and 10, in model (15),

$$r_{\mathrm{np}} = O_p(pn^{-1/2}),$$
$$r_{\mathrm{inf}} = O_p(n^{-1/2}).$$

5.3. Bayesian asymptotics

We briefly discuss the Bayesian version of r^* that is obtained from the Laplace approximation (Reid (2003), section 2.2). Given a prior density $\pi(\psi, \lambda)$ on the parameter space, the tail area of the marginal posterior distribution for ψ is approximated by $\Phi(r_{\mathrm{B}}^*)$, where

$$r_{\mathrm{B}}^* = r + \frac{1}{r}\log\left(\frac{q_{\mathrm{B}}}{r}\right),$$

and

$$q_{\mathrm{B}} = s\rho^{-1}\frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_{\psi_0})}.$$

We write

$$r_{\mathrm{B}}^* = r + r_{\mathrm{np}} + r_{\mathrm{inf}} + r_{\mathrm{prior}},$$

where r_{np} and r_{inf} are the same as in the location–scale model, and $r_{\mathrm{prior}} = r^{-1}\log\{\pi(\hat{\theta})/\pi(\hat{\theta}_{\psi_0})\}$.

Proposition 2 is valid under the same assumptions as given in Section 5.2. However, the con-

tribution of the prior can be non-negligible, unlike the p-fixed asymptotic regime. For example in the simplest case where the priors for each parameter are independent,

$$r_{\text{prior}} = \frac{1}{r}\left[\log\left\{\frac{\pi_1(\hat{\psi})}{\pi_1(\psi_0)}\right\} + \sum_{j=2}^{p} \log\left\{\frac{\pi_j(\hat{\theta}_j)}{\pi_j(\hat{\theta}_{j,\psi_0})}\right\}\right]$$
$$= \frac{1}{r}\left[\frac{\partial}{\partial\psi}\log\{\pi_1(\psi)\}|_{\psi=\tilde{\psi}}(\psi_0-\hat{\psi}) + \sum_{j=1}^{p}\frac{\partial\hat{\theta}_{j,\psi}}{\partial\psi}\bigg|_{\psi=\tilde{\psi}}\frac{\partial}{\partial\theta_j}\log\{\pi_j(\theta_j)\}|_{\theta_j=\hat{\theta}_{\tilde{\psi},j}}(\psi_0-\hat{\psi})\right],$$

where π_j denotes the prior for the jth parameter θ_j, and $\tilde{\psi}$ and $\hat{\theta}_{\tilde{\psi},j}$ are as defined in Section 2.1. Using the same set of assumptions as in proposition 2, we have $\psi_0 - \hat{\psi} = O_p(n^{-1/2})$ and the derivative of the constrained maximum likelihood is $O_p(n^{-1/2})$ by lemma 1 in the on-line supplementary materials. Further assuming that the derivatives of the log-prior-density are uniformly bounded in a Euclidean ball of radius δ around $\theta_{0,j}$, where $\theta_{0,j}$ denotes the jth component of θ_0, this results in $r_{\text{prior}} = O_p(n^{-1/2})$ under our assumption of $p = o(n^{1/2})$. This shows that under our assumptions the effect of the prior is the same as in the usual p-fixed asymptotics.

6. Simulations

6.1. Example: logistic regression

The model is

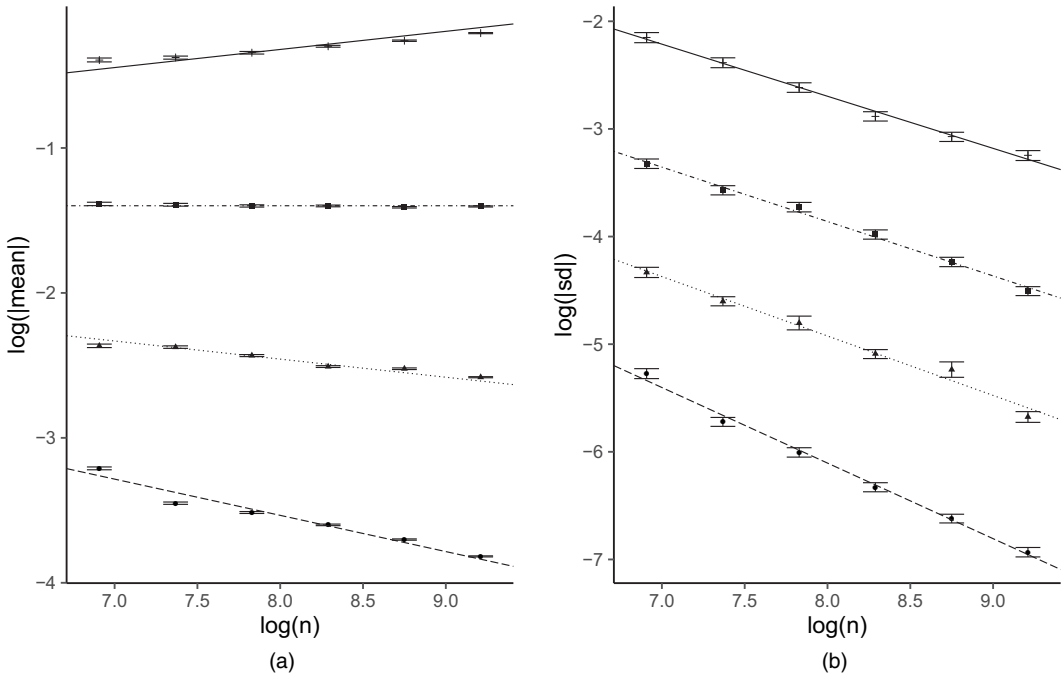

Fig. 1. Behaviour of r_{np} for logistic regression, determined from 1000 simulations: (a) comparison of the logarithm of the empirical mean with the line of slope $\alpha - \frac{1}{2}$ (we see that the log-means are close to the line); (b) simulated standard deviation on the log-scale, and the regression line through the six points (the fitted slopes for the log-standard-deviation are all around $-\frac{1}{2}$ or smaller; from this we can see that r_{np} is mainly adjusting for a location bias in r, as the values plotted in (a) are an order of magnitude larger than the values plotted in (b)) (●, scaling 0.25; ▲, scaling 0.375; ■, scaling 0.5; +, scaling 0.625)

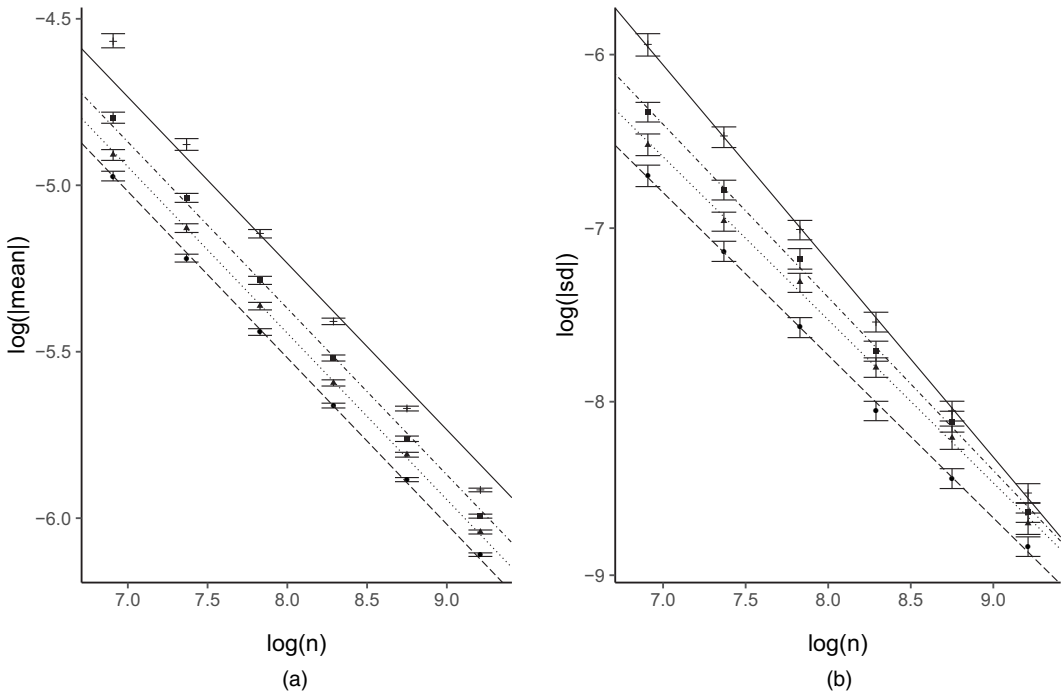

Fig. 2. Plots illustrating the behaviour of r_{inf} for logistic regression, determined from 1000 simulations: (a) comparison of the logarithm of the empirical mean with the line of slope $-\frac{1}{2}$ (we see that the log-means are close to the line); (b) simulated standard deviation on the log-scale, and the regression line through the six points (the slopes of the fitted lines are smaller than $-\frac{1}{2}$; this suggests that the location adjustment is dominant as the values plotted in (a) are an order of magnitude larger than the values plotted in (b)) (●, scaling 0.25; ▲, scaling 0.375; ■, scaling 0.5; +, scaling 0.625)

$$y_i \sim \text{Bern}(p_i), \qquad p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1+\exp(x_i^{\mathrm{T}}\beta)}.$$

We generated n vectors x_i of length p from a multivariate normal distribution with $\mathbb{E}(x_{ij}) = 0$, $\text{var}(x_{ij}) = 1$ and $\text{cov}(x_{ij}, x_{ik}) = 0.9^{|j-k|}$. This covariance structure was chosen so that the maximal and minimal eigenvalues of the covariance matrix are bounded above and below, and the correlation between x_{ij} and x_{ik} is non-zero. The true values of the regression coefficients were taken as $\beta_0 = \beta_1 = 1$ and $\beta_i = 1/\sqrt{p}$ for $i = 2, \ldots, p$. The parameter of interest is β_1.

For each combination of n and p we simulated 1000 values of r_{inf} and r_{np}, and computed p-values for testing $H_0 : \beta_1 = 1$ based on the normal approximation to the distribution of r and of r^*. We used the sets of values $n = \{10^3, 10^{3.2}, 10^{3.4}, 10^{3.6}, 10^{3.8}, 10^4\}$ and $p = \{n^{0.25}, n^{0.375}, n^{0.5}, n^{0.625}\}$. If a random variable $Z_n = O_p(n^\nu)$, then we expect one or both of $|\mathbb{E}(Z_n)|$ and $\text{var}(Z_n)^{1/2}$ to be linear in $\log(n)$ with slope ν. As $p = O(n^\alpha)$ for $0 < \alpha < \frac{1}{2}$, according to our theoretical results $r_{\text{np}} = O_p(p/n^{1/2})$ and $r_{\text{inf}} = O_p(n^{-1/2})$, we would expect either or both of the log-expectation and the log-standard-deviation of r_{np} to have slope $\alpha - \frac{1}{2}$, and either or both of the log-expectation and the log-standard-deviation of r_{inf} to have a slope of $-\frac{1}{2}$.

In Figs 1 and 2 we plot the 95% bootstrapped confidence intervals of the empirical mean and standard deviation from 1000 simulations as a function of n, on the log-scale. As mentioned in Section 2, small values of r may produce numerical instabilities producing

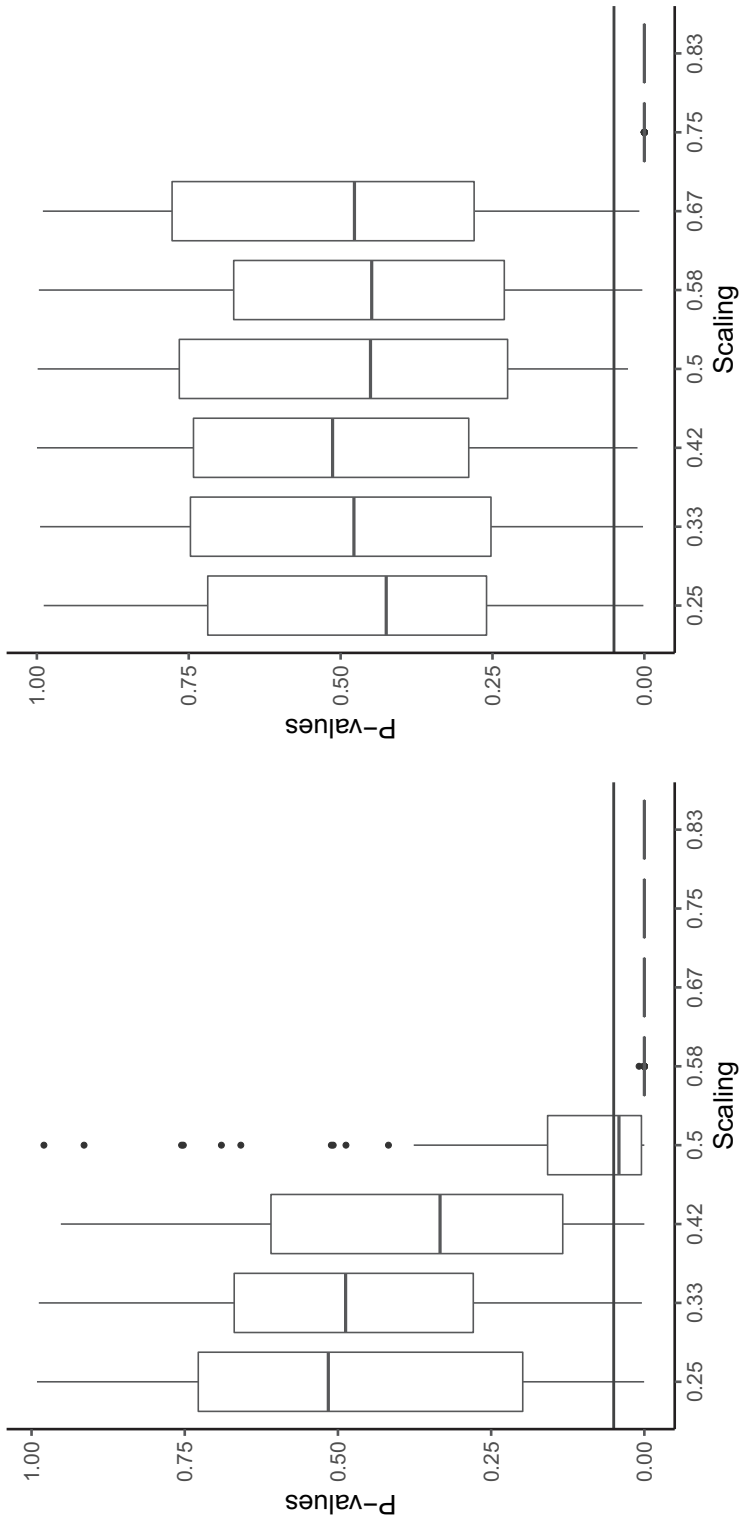

Fig. 3. Plots for logistic regression illustrating the difference in the breakdown point of uniformity of the p-value distribution based on the standard normal approximation to the distribution of (a) r and of (b) r^*: we see that p-values based on the r^*-approximation appear to be uniformly distributed up to about $p = O(n^{2/3})$, whereas those based on the normal approximation to the distribution of r begin to exhibit non-uniformity at about $p = O(n^{1/2})$

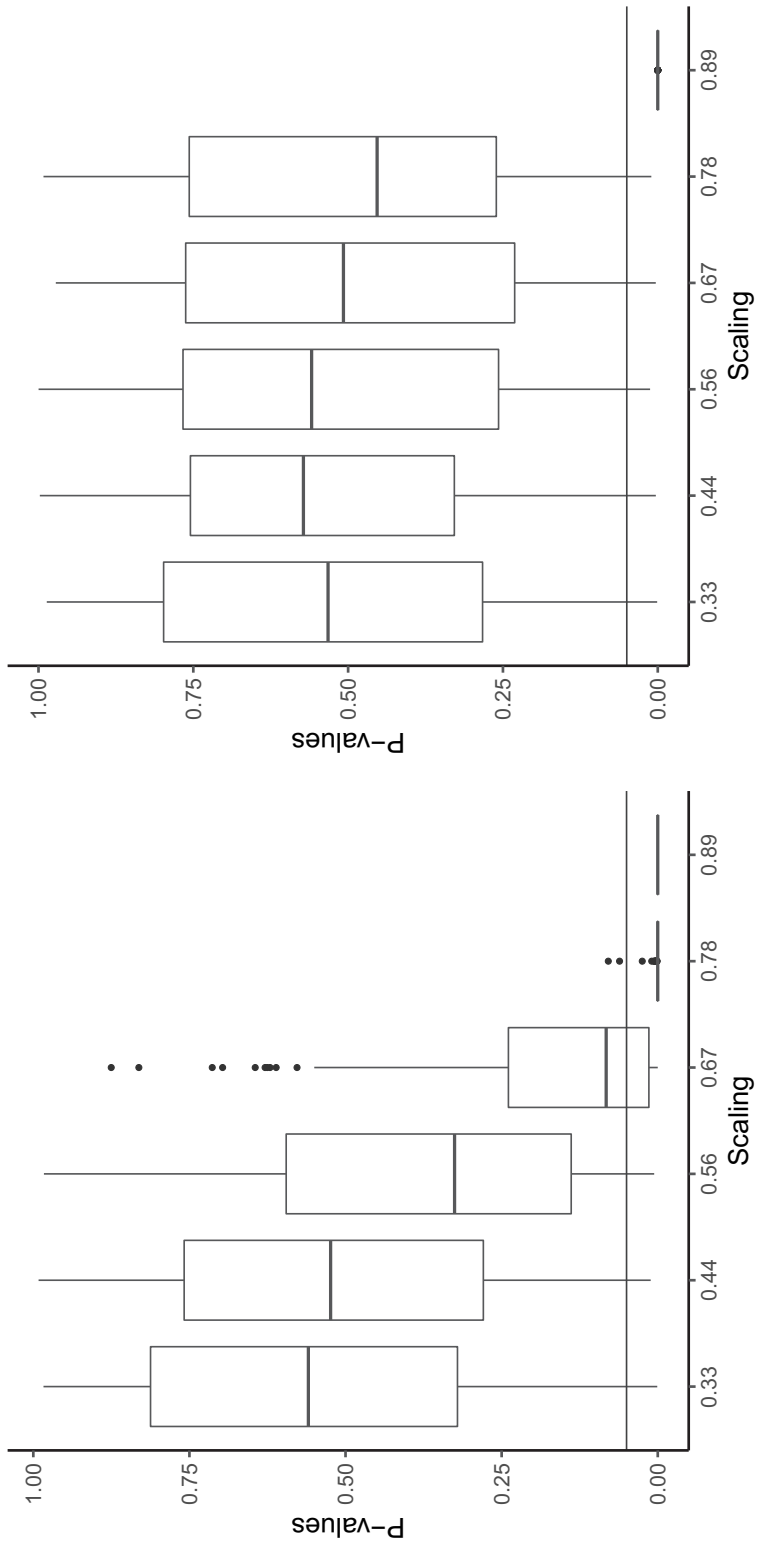

Fig. 4. Plots for logistic regression illustrating the difference in the breakdown point of uniformity of the p-value distribution for p-values based on the normal approximation to the distribution of (a) r and of (b) r^* under the global null

unreliable estimates for the standard deviation and mean, so we consider only samples with $|r| > 0.025$. The slopes of the lines in Figs 1 and 2 are approximately $\alpha - \frac{1}{2}$ for the expectation of r_{np} and $-\frac{1}{2}$ for the expectation of r_{inf} and generally less than $-\frac{1}{2}$ for the standard deviation of r_{inf} and r_{np}, which is consistent with our theoretical results.

To assess the null distribution, we examine the uniformity of the simulated p-values. We fix $n = 1000$ and $p = n^{\alpha}$ with $\alpha = (3/12, 4/12, 5/12, 6/12, 7/12, 8/12, 9/12, 10/12)$. The simulation settings are as above. For each value of p we obtained 1000 simulated p-values by using the standard normal approximation to the distribution of r and of r^*. We tested the assumption that these simulation p-values were distributed as $U(0, 1)$ by using the Kolmogorov–Smirnov test. This was repeated 100 times, giving 100 p-values from a Kolmogorov–Smirnov test for uniformity. Boxplots of the p-values of these uniformity tests for various values of α are displayed in Fig. 3. In Fig. 3(a) we see that p-values based on the normal approximation to the distribution of r exhibit non-uniformity around $p = n^{1/2}$; in Fig. 3(b) p-values based on the normal approximation to the distribution of r^* are consistent with $U(0, 1)$ up to a scaling of roughly $p = n^{2/3}$.

We also examine the breakdown point supposing that we are under the global null where $\beta_i = 0$ for $i = 0, \ldots, p$. It was theoretically determined in Fan *et al.* (2019) that with Gaussian covariates the exact breakdown point for the Wald test for a single parameter is $p = n^{2/3}$. Our simulations show that the same breakdown point holds for r. For this example we choose a more aggressive set of scalings $\alpha = (3/9, 4/9, 5/9, 6/9, 7/9, 8/9)$. In Fig. 4 we see that once again p-values based on the normal approximation to the distribution of r^* are approximately uniformly distributed up to a scaling of roughly $p = O(n^{7/9})$, whereas those based on the normal approximation to the distribution of r break down at roughly $p = O(n^{2/3})$.

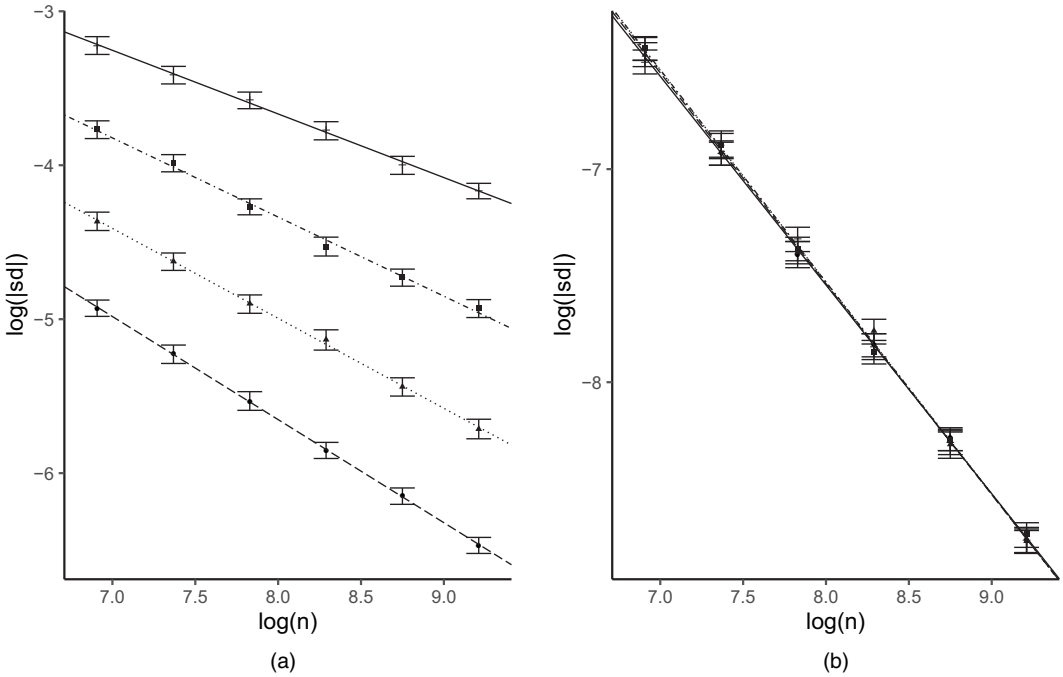

Fig. 5. Plots for Weibull regression illustrating the magnitude of r_{inf} and r_{np} (we plot the log-standard-deviation of r_{np} and r_{inf}, with the line of best fit for both plots): (a) the fitted slopes for r_{np} are less than $\alpha - \frac{1}{2}$ (●, scaling 0.33; ▲, scaling 0.42; ■, scaling 0.5; +, scaling 0.58); (b) the fitted slopes for the log-standard-deviation of r_{inf} are much less than $-\frac{1}{2}$ (●, scaling 0.33; ▲, scaling 0.42; ■, scaling 0.5; +, scaling 0.58)

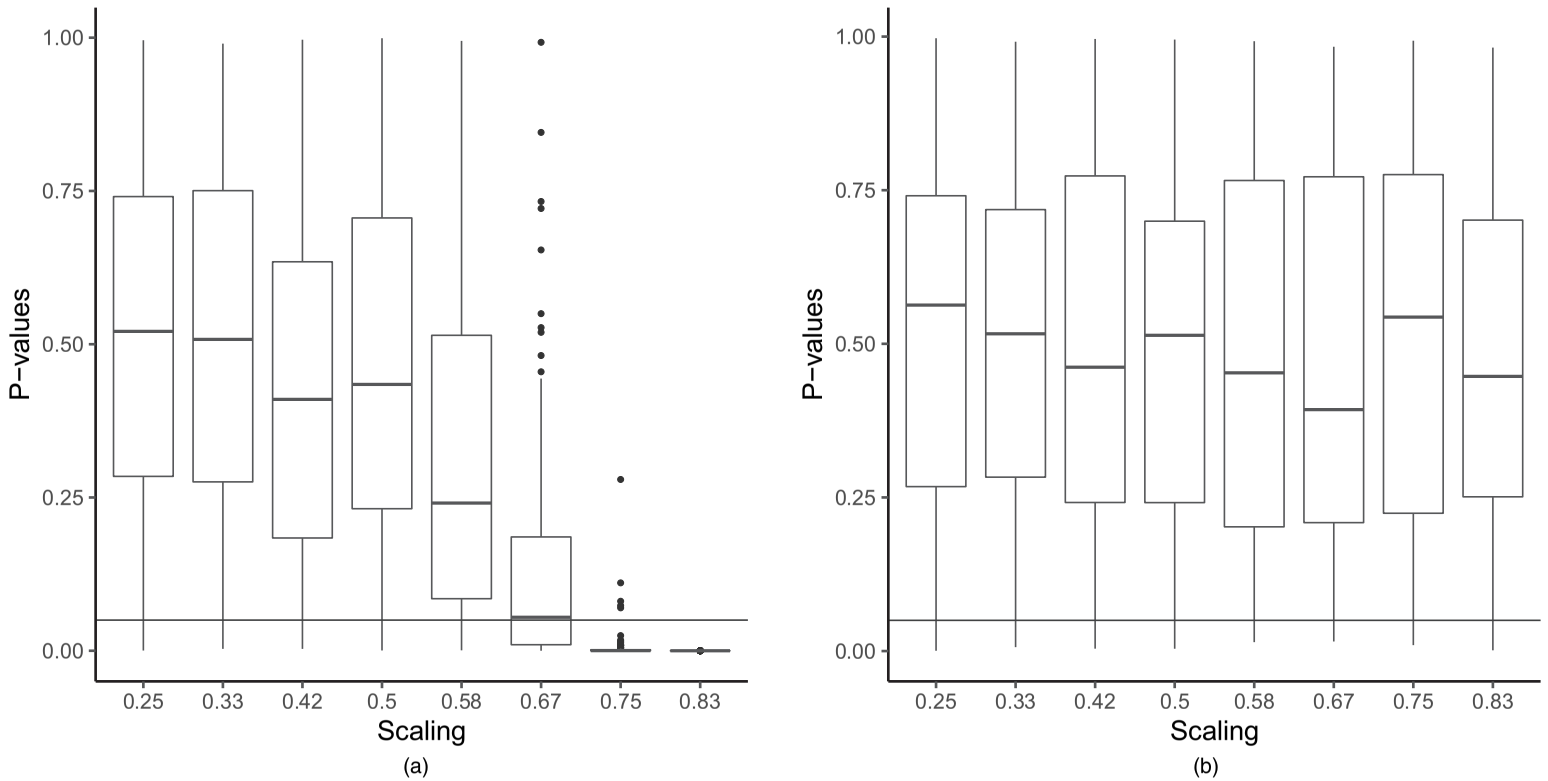

Fig. 6. Plots for Weibull regression illustrating the difference in the breakdown point of uniformity of the p-value distribution for (a) r and (b) r^*: we see that r^* still maintains uniformity for all scaling, whereas r begins to exhibit non-uniformity around $p = O(n^{2/3})$

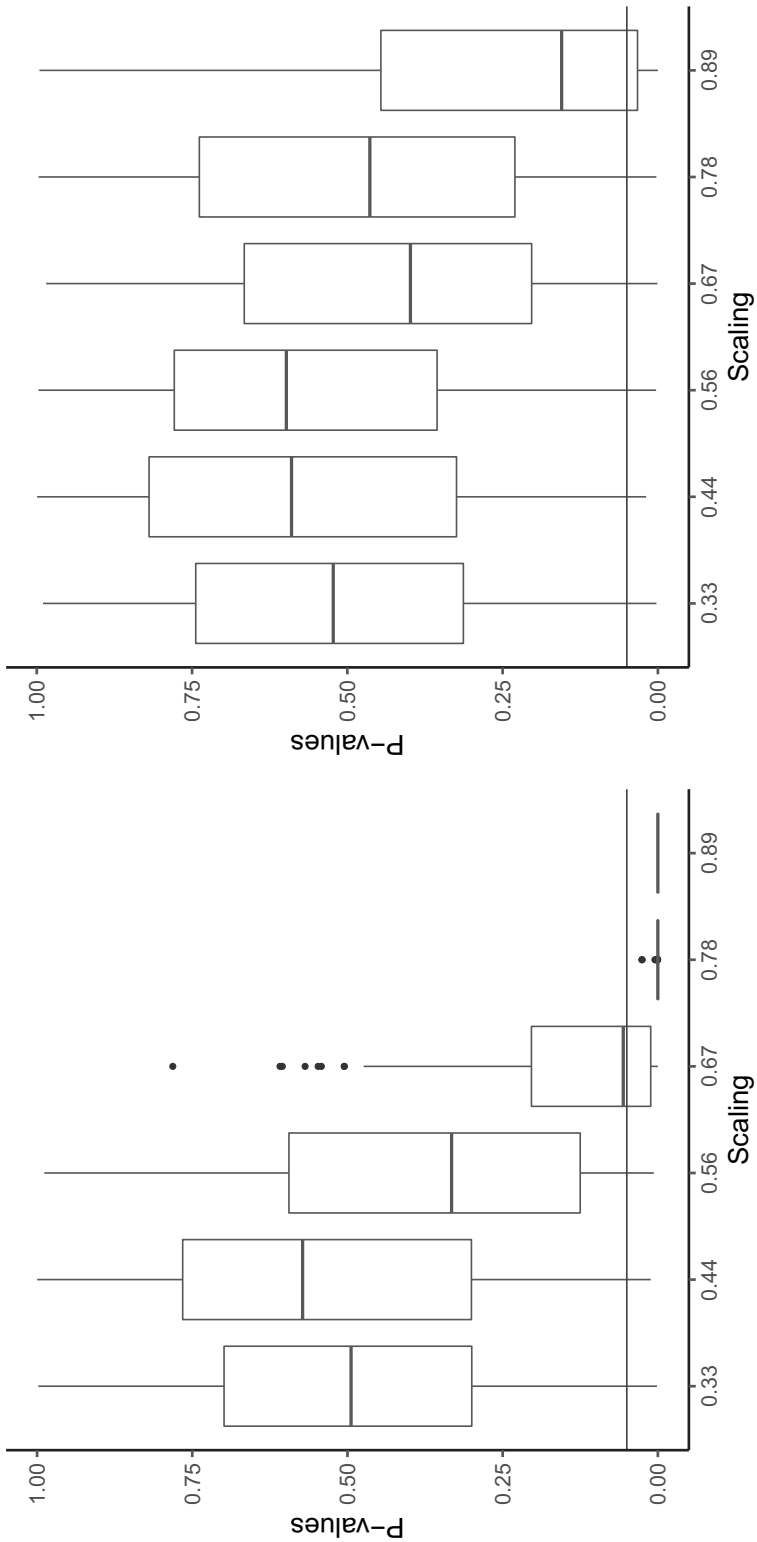

Fig. 7. Plots for Weibull regression illustrating the difference in the breakdown point of uniformity of the p-value distribution based on the normal approximation to the distributions of (a) r and (b) r^*: we see that the r^*-approximation is consistent with uniformity to $\alpha = 8/9$, whereas for r we see evidence of non-uniformity around $p = O(n^{2/3})$

6.2. Example: Weibull regression

As an illustration of the location–scale model we simulate observations from a Weibull regression:

$$y_i = x_i^{\mathrm{T}}\beta + \sigma z_i,$$

where $f(z) = \exp\{z - \exp(z)\}$. We generated 1000 simulations from this model with regression coefficients $\beta_0 = 1$, $\beta_1 = 0$, $\beta_i = 1/\sqrt{p}$ for $i = 2, \ldots, p$ and $\sigma = 2$. We use six possible values of $n = \{10^3, 10^{3.2}, 10^{3.4}, 10^{3.6}, 10^{3.8}, 10^4\}$, and $p = \{n^{4/12}, n^{5/12}, n^{6/12}, n^{7/12}\}$. For each simulation we obtained the p-value for testing $H_0 : \beta_1 = 0$ based on the normal approximation to the distribution of r and of r^*. We note that, although it is possible to orthogonalize β_i to β_1 for $i = \{2, \ldots, p\}$, σ is not orthogonal to β_1 as the density of the Weibull regression is not symmetric. Therefore, it is not obvious whether assumption 2 holds; nevertheless our results seem to be valid for the Weibull regression in the simulated results.

Plots of the simulated standard deviation of r_{np} and r_{inf} are given in Fig. 5. Again we consider only samples with $|r| > 0.025$ to avoid the singularity near $\hat{\psi} = \psi_0$. The empirical means were essentially 0, so they are not plotted here. We estimated the standard deviation empirically from the 1000 generated values and plot the 95% bootstrap confidence intervals against their associated value of $\log(n)$. The slope of the line for the standard deviation for each of the scalings of p illustrates the order in n of r_{inf} and r_{np}. Fig. 5 shows that the theoretical prediction for r_{inf} is correct; however, it appears that the scaling for r_{np} is slower than expected. On the basis of results in Section 5.2 we would have expected slopes of $\alpha - \frac{1}{2}$; however, it appears that the slopes obtained are smaller. This demonstrates that the scaling rates of r_{np} are better than our conservative upper bounds for some location–scale models.

As in Section 5.1, we assessed the uniformity of the p-values based on the normal approximation to the distribution of r and r^*. We fixed $n = 1000$ and considered various possible scalings of $p = n^{\alpha}$ with $\alpha = \{0.25, 0.33, 0.42, 0.50, 0.58, 0.67, 0.75, 0.83\}$. As above the regression coefficients were set to $\beta_0 = 1, \beta_1 = 0$ and $\beta_i = 1/\sqrt{p}$ for $i = 2, \ldots, p$, and we tested $H_0 : \beta_1 = 0$.

The results are displayed in Fig. 6, where it is apparent that the normal approximation to the distribution of r^* is much more accurate than the normal approximation to the distribution of r. We see that p-values based on r exhibit non-uniformity around $p = n^{2/3}$ whereas those based on r^* maintain the uniformity of the distribution of the p-values for all scalings displayed. This is quite remarkable as, for $\alpha = 0.83$, $p = 464$, meaning that the number of covariates is almost half the number of observations.

We also examined the breakdown point of the p-value distribution under the global null $\beta_i = 0$ for $i = 0, \ldots, p$. For this example we chose a different set of possible scalings $\alpha = \{0.33, 0.44, 0.56, 0.67, 0.78, 0.89\}$. We used the same procedure to simulate x_i as described above. In Fig. 7 we see again that p-values based on the r^*-approximation are uniformly distributed under the null up to a higher scaling; the normal approximation to the distribution of r still breaks down at $p = O(n^{2/3})$, whereas that for r^* breaks down around $p = O(n^{8/9})$.

7. Discussion

Theorems 1 and 2 in Section 3 establish the size of the two correction terms for the likelihood root, as a function of the dimension p and the sample size n, although the numerical work in Section 6 shows that in special cases the rate may be better than what is proved. The results also provide an explanation for the observation that correction for nuisance parameters is the most important aspect of higher order approximations.

These results suggest that higher order approximations can be broadly applied in models where the number of nuisance parameters is comparable with the number of observations and are not only improvements in small sample settings.

The expansions that were used in the proofs of results in Section 3 rely on an intermediate value of ψ; an alternative approach would be to continue the Taylor series as a formal asymptotic expansion. The approach that was used here makes it easier to control the error term. The formal approach is explored in a companion paper (Tang and Reid, 2020), as it provides some insight into the structure of the expressions in models with many nuisance parameters.

Research directions that could be explored building on this work include the following suggestions:

(a) verifying that the normal approximation to the distribution of r^* is more accurate than the normal approximation to the distribution of r in the high dimensional setting, as is suggested by the simulations;
(b) analysing inference based on the modified profile likelihood function, as Sartori (2003) showed that in the stratified model setting this has better asymptotic behaviour as p increases than inference based on r^*;
(c) using the techniques that were developed here to study related higher order approximations, including for example the p^*-approximation to the density of the maximum likelihood estimator, and saddlepoint approximations to the density of M-estimators. This may be accomplished by extending the work of Field (1982) to the high dimensional regime, and the broader generality of M-estimates will be helpful in studying the behaviour of estimators that are not likelihood based.
(d) A reviewer asked whether the constrained maximum likelihood estimator could be replaced by a nuisance parameter estimator that exploits sparsity, as is often used in models with many parameters. We think that this would raise several technical difficulties, but analysis of the M-estimator that was described above might be helpful for this.

Acknowledgements

We thank Nicola Sartori, Michele Lambardi di San Miniato, Ioannis Kosmidis, Heather Battey and Michaël Lalancette for helpful discussions. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada and the Vector Institute.

References

Barndorff-Nielsen, O. E. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.
Barndorff-Nielsen, O. E. and Cox, D. R. (1994) *Inference and Asymptotics*. London: Chapman and Hall.
Bartlett, M. (1937) Properties of sufficiency and statistical tests. *Proc. R. Soc.* A, **160**, 268–282.
Bartlett, M. (1953) Approximate confidence intervals ii. *Biometrika*, **40**, 306–317.
Brazzale, A., Davison, A. and Reid, N. (2007) *Applied Asymptotics: Case Studies in Small-sample Statistics*. Cambridge: Cambridge University Press.
Cox, D. (1988) Some aspects of conditional and asymptotic inference. *Sankhya* A, **50**, 314–337.
Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B, **49**, 1–39.
Cox, D. R. and Reid, N. (1992) A note on the difference between profile and modified profile likelihood. *Biometrika*, **79**, 408–411.
Davison, A. C. (1988) Approximate conditional inference in generalized linear models. *J. R. Statist. Soc.* B, **50**, 445–461.
Fan, Y., Demirkaya, E. and Lv, J. (2019) Nonuniformity of p-values can occur early in diverging dimensions. *J. Mach. Learn. Res.*, **20**, 1–33.
Field, C. (1982) Small sample asymptotic expansions for multivariate M-estimates. *Ann. Statist.*, **10**, 672–689.

Kosmidis, I., Kenne Pagui, E. C. and Sartori, N. (2019) Mean and median bias reduction in generalized linear models. *Statist. Comput.*, **30**, 43–59.
Lawley, D. (1956) A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, **43**, 295–303.
Lei, L., Bickel, P. J. and Karoui, N. E. (2016) Asymptotics for high dimensional regression M-estimates: fixed design results. *Probab. Theory Reltd Flds*, **172**, 983–1079.
McCullagh, P. and Tibshirani, R. (1990) A simple method for the adjustment of profile likelihoods. *J. R. Statist. Soc.* B, **52**, 325–344.
Pierce, D. A. and Peters, D. (1992) Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. R. Statist. Soc.* B, **54**, 701–737.
Portnoy, S. (1984) Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large: i, consistency. *Ann. Statist.*, **12**, 1298–1309.
Portnoy, S. (1988) Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.*, **16**, 356–366.
Reid, N. (2003) Asymptotics and the theory of inference. *Ann. Statist.*, **31**, 1695–1731.
Sartori, N. (2003) Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*, **90**, 533–549.
Sartori, N., Bellio, R., Salvan, A. and Pace, L. (1999) The directed modified profile likelihood in models with many nuisance parameters. *Biometrika*, **86**, 735–742.
Shun, Z. and McCullagh, P. (1995) Laplace approximation of high dimensional integrals. *J. R. Statist. Soc.* B, **57**, 749–760.
Sur, P. and Candès, E. J. (2019) A modern maximum likelihood theory for high-dimensional logistic regression. *Proc. Natn. Acad. Sci. USA*, **116**, 14516–14525.
Sur, P., Chen, Y. and Candès, E. J. (2019) The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probab. Theory Reltd Flds*, **175**, 487–558.
Tang, Y. and Reid, N. (2020) Modified likelihood root as a polynomial of the likelihood root. *Manuscript*. Department of Statistical Science, University of Toronto, Toronto.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for: Modified likelihood root in high dimensions'.