OXFORD UNIVERSITY PRESS | Biometrika

Accurate directional inference for vector parameters

Journal:	*Biometrika*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Fraser, Don; University of Toronto, Department of Statistics Reid, Nancy; University of Toronto, Department of Statistics Sartori, Nicola; University of Padova, Department of Statistical Sciences
Keywords:	Behrens-Fisher problem, Box-Cox model, Higher-order Asymptotics, Likelihood Ratio Test, Marginal Independence, Tangent Exponential Model

Biometrika (2012), **99**, 1, *pp*. 1–16

Printed in Great Britain

Accurate directional inference for vector parameters

BY D. A. S. FRASER, N. REID
Department of Statistics, University of Toronto, Toronto, Canada M5S 3G3
dfraser@utstat.toronto.edu reid@utstat.toronto.edu

AND N. SARTORI
Dipartimento di Scienze Statistiche, Università degli Studi di Padova, 35121 Padova, Italy
sartori@stat.unipd.it

SUMMARY

We consider inference on a vector-valued parameter of interest in a parametric model, in the presence of a finite-dimensional nuisance parameter. Based on higher order asymptotic theory for likelihood, we propose a directional test whose p-value is computed using one-dimensional integration. This extends the development of Davison et al. (2014) from linear exponential families to more general families of models. Examples and simulations illustrate the high accuracy of the method, which we compare with the usual likelihood ratio test and with an adjusted version of it. In high-dimensional settings the approach works essentially perfectly, whereas its competitors can fail catastrophically.

Some key words: Behrens-Fisher problem; Box-Cox model; Higher-order Asymptotics; Likelihood Ratio Test; Marginal Independence; Tangent Exponential Model.

1. INTRODUCTION

In this paper we consider likelihood inference for vector parameters in parametric models, generalizing the directional approach of Davison et al. (2014). That paper restricted attention to parametric models that were exponential families, and to parameters of interest that were linear in the canonical parameter of the exponential family model. Here we extend the method to non-linear parameters in exponential families, and to non-exponential family models. For the latter the starting point is the tangent exponential model (Fraser & Reid, 1995; Brazzale et al., 2007, Ch. 8).

The main likelihood-based method of inference for vector parameters is the log-likelihood ratio statistic, which is typically referred to a χ^2 distribution, with degrees of freedom equal to the number of parameters of interest. The χ^2 approximation can be improved by either a large-deviation method proposed in Skovgaard (2001), or by Bartlett correction (Bartlett, 1937); the latter rescales the likelihood ratio statistic by its expected value under the hypothesis that fixes the parameter of interest. The likelihood ratio statistic and these two adjusted versions are omnibus test statistics, in the sense that values 'as large or larger' than that observed can correspond to any part of the parameter space of the vector parameter of interest.

The directional method, originally proposed in Fraser & Massam (1985), and developed further in Skovgaard (1988), adapts the test statistic to the direction of departure from the hypothesis indicated by the observed data point. The p-value thus obtained is analogous to a two-sided p-value obtained by doubling the smaller of two one-sided p-values. In Davison et al. (2014), it

was applied to inference for contingency tables, logistic regression, equality of normal variances, and Gaussian graphical models. Simulations presented there indicated that the method was very effective at adjusting for the estimation of possibly large numbers of nuisance parameters, as well as maintaining the size of the test at very nearly the nominal value. In contrast, the χ^2 approximation to the likelihood ratio test, and its corrected versions, were in some settings very inaccurate.

One advantage of linear exponential families is that there is a well-defined conditional distribution, free of the nuisance parameters, on which inference for the parameters of interest can be based. This provides an 'exact' comparison in simple problems, and also simplifies the derivation of the directional test. It turns out that the method however can be extended to more general families, with a slight increase in the complexity of the formulas. We give some background in §2, present the directional test in §3, and illustrate it on three problems in §4: the Behrens-Fisher problem, marginal independence, and the Box-Cox model. As in the exponential family case with linear parameter of interest, the simulation results indicate that the method is remarkably accurate, and more reliable than either the likelihood ratio test or Skovgaard's (Skovgaard, 2001) large deviation method, especially for high-dimensional parameters.

2. LIKELIHOOD RATIO TESTS

We assume a parametric model $f(y;\theta)$ for a vector of independent components $y = (y_1, \ldots, y_n)$, with $\theta \in \mathbb{R}^p$. The log-likelihood function is $\ell(\theta) = \ell(\theta; y) = \log f(y;\theta)$, and the maximum likelihood estimator $\widehat{\theta} = \widehat{\theta}(y)$ maximizes the log-likelihood function. When needed we denote the observed data point by y^0, with associated log-likelihood $\ell^0(\theta) = \ell(\theta; y^0)$ and maximum likelihood estimate $\widehat{\theta}^0 = \widehat{\theta}(y^0)$.

We suppose that we have a d-dimensional parameter of interest, $\psi(\theta)$, and denote by H_ψ the hypothesis $\psi(\theta) = \psi$. It may often be the case that ψ is a component of the full parameter θ, possibly after re-parameterization. We let $\widehat{\theta}_\psi$ denote the constrained maximum likelihood estimator of θ under H_ψ; if $\theta = (\psi, \lambda)$, then $\widehat{\theta}_\psi = (\psi, \widehat{\lambda}_\psi)$.

Under standard regularity conditions, the limiting distribution of $\widehat{\theta}$ suitably centered and scaled, is standard normal. Thus its distribution in finite samples can be approximated by a normal distribution with mean θ and covariance matrix estimated by $\jmath^{-1}(\widehat{\theta})$, where $\jmath(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^\top$ is the observed Fisher information function; an analogous result holds for $\widehat{\theta}_\psi$ under H_ψ (Cox & Hinkley, 1974, §9.3). A parameterization-invariant measure of departure of $\widehat{\theta}$ from H_ψ is given by the likelihood ratio statistic

$$w(\psi) = 2\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\psi)\}. \tag{1}$$

The limiting distribution of $w(\psi)$ is χ^2_d, where d is the number of constrained parameters in H_ψ.

Skovgaard (2001) proposed an improvement of the likelihood ratio test based on arguments developed for large-deviation approximations in the case where the parameter of interest is scalar. His adjusted version of $w(\psi)$ is

$$w^*(\psi) = w\left(1 - \frac{\log\gamma}{w}\right)^2, \tag{2}$$

to be referred to a χ^2_d distribution: the correction factor γ compares w to an asymptotically equivalent quadratic form. Another improved approximation to the distribution of $w(\psi)$ can be

obtained by Bartlett correction (Bartlett, 1937):

$$w_B(\psi) = \frac{w(\psi)}{\mathrm{E}_{\widehat{\theta}_\psi}\{w(\psi)\}}, \tag{3}$$

where in principle the expected value of the likelihood ratio statistic can either be computed analytically, by an asymptotic expansion, or by simulation. The χ^2_d approximation to the distribution of $w_B(\psi)$ has relative error $O(n^{-2})$ (Barndorff-Nielsen & Cox, 1994), whereas that for $w(\psi)$ is $O(n^{-1})$. Skovgaard (2001) argued that $w^*(\psi)$ is easier to calculate than $w_B(\psi)$, and also noted that estimating $\mathrm{E}_{\widehat{\theta}_\psi}\{w(\psi)\}$ from its asymptotic expansion did not give very accurate finite sample approximations.

Tests based on $w(\psi)$, $w^*(\psi)$ and $w_B(\psi)$ give omnibus measures of departure; all potential directions away from the hypothesis H_ψ are averaged in the calculation of the p-value. In the next section we propose a measure of departure that incorporates information in the data about the relevant direction of deviation from H_ψ, by conditioning, generalizing the analogous measure for linear exponential families of Davison et al. (2014).

3. Directional tests

If the density for y is an exponential family model:

$$f(y;\theta) = \exp[\varphi(\theta)^\top u(y) - K\{\varphi(\theta)\}]f_0(y), \tag{4}$$

where $\varphi(\theta)$ is the canonical parameterization, then the log-likelihood function is

$$\ell(\theta;u) = \varphi(\theta)^\top u - K\{\varphi(\theta)\} = \varphi(\theta)^\top(u-u^0) + \ell(\theta;u^0), \tag{5}$$

which we express in terms of the centered version of the sufficient statistic, $s = u - u^0$:

$$\ell(\theta;s) = \varphi(\theta)^\top s + \ell(\theta;s^0=0) = \varphi(\theta)^\top s + \ell^0(\theta). \tag{6}$$

If the density for y is not in the exponential family, we approximate its density using the tangent exponential model (Fraser & Reid, 1995; Brazzale et al., 2007), and the log-likelihood function of this model has the same form (6). The tangent exponential family model is constructed as an approximation to the conditional distribution of y given a locally defined ancillary statistic. It agrees to first derivative with the original model at the observed data point y^0, and having an exponential family form, has a corresponding saddlepoint approximation which can be used as a basis for calculations giving approximations in continuous models to $O(n^{-3/2})$. Some further detail is given in the Appendix.

The canonical parameter in (6), $\varphi = \varphi(\theta)$ is a reparameterization of $\theta = (\psi, \lambda)$, and we assume that the parameter of interest ψ is a nonlinear function of φ. The directional p-value uses as its starting point the saddlepoint approximation to the density of s determined by (6), constrained to the d-dimensional sample space $\mathcal{L}_\psi$ defined by fixing the constrained maximum likelihood estimator under the hypothesis at its observed value $\widehat{\varphi}^0_\psi$. This approximation is

$$h(s;\psi) = c\exp\left\{\ell(\widehat{\varphi}^0_\psi;s) - \ell(\widehat{\varphi}(s);s)\right\}\left|\jmath_{\varphi\varphi}(\widehat{\varphi}(s);s)\right|^{-1/2}\left|\jmath_{(\lambda\lambda)}(\widehat{\varphi}^0_\psi;s)\right|^{1/2},\ s\in\mathcal{L}_\psi, \tag{7}$$

where $\ell(\varphi;s)=\varphi^\top s+\ell^0\{\theta(\varphi)\}$, as at (6), $\widehat{\varphi}(s)$ is the corresponding maximum likelihood estimate, obtained by solving $\ell_\varphi(\widehat{\varphi}(s);s)=0$,

$$|\jmath_{\varphi\varphi}(\varphi;s)|=\Big|-\frac{\partial^2\ell(\varphi;s)}{\partial\varphi\partial\varphi^\top}\Big|=\Big|-\frac{\partial^2\ell^0(\theta(\varphi))}{\partial\varphi\partial\varphi^\top}\Big|=|\jmath\{\varphi(\theta)\}|\,|\varphi_\theta(\theta)|^{-2}\,, \tag{8}$$

$$|\jmath_{(\lambda\lambda)}(\varphi;s)|=|\jmath_{\lambda\lambda}(\varphi;s)|\,|\varphi_\lambda(\theta)^\top\varphi_\lambda(\theta)|^{-1}\,, \tag{9}$$

with $\jmath_{\lambda\lambda}(\varphi;s)=-\partial^2\ell(\varphi;s)/\partial\lambda\partial\lambda^\top$, $\varphi_\theta(\theta)=\partial\varphi(\theta)/\partial\theta$, and $\varphi_\lambda(\theta)=\partial\varphi(\theta)/\partial\lambda$. The density h for s has a d-dimensional parameter, and a d-dimensional sample space, $\mathcal{L}_\psi$; the nuisance parameter λ has been eliminated by marginalization, as sketched in the Appendix.

We measure the departure of s from its observed value $s^0=0$, along the line in $\mathcal{L}_\psi$, say $\mathcal{L}_\psi^*$, that joins s^0 with s_ψ, which is the value of s leading to the observed constrained maximum likelihood estimate ($\widehat{\varphi}(s_\psi)=\widehat{\varphi}^0_\psi$):

$$s_\psi+\ell^0_\varphi(\widehat{\varphi}^0_\psi)=0. \tag{10}$$

The line from s_ψ through s^0 is parameterized by $t\in\mathbb{R}$,

$$s(t)=s_\psi+t(s^0-s_\psi)=(1-t)s_\psi\,, \tag{11}$$

the second form following as $s^0=0$.

As t increases, $\widehat{\varphi}\{s(t)\}$ traces out a curve in the parameter space that passes through the constrained maximum likelihood estimate $\widehat{\varphi}^0_\psi$ when $t=0$ and through the full maximum likelihood estimate $\widehat{\varphi}^0$ when $t=1$. Because $s=s(t)$, is constrained to $\mathcal{L}_\psi$, the second determinant on the right hand side of (9) does not depend on t and therefore can be ignored in the computation of the directional p-value. Moreover, if ψ is a linear function of the natural parameter φ of an exponential family, as in Davison et al. (2014), even the first factor on the right hand side of (9) does not depend on t and then the entire last factor of (7) can be omitted.

The directional p-value measuring the departure from H_ψ on the line $\mathcal{L}_\psi^*$ is then

$$p(\psi)=\frac{\int_1^{t_{\max}}t^{d-1}h\{s(t);\psi\}\,\mathrm{d}t}{\int_0^{t_{\max}}t^{d-1}h\{s(t);\psi\}\,\mathrm{d}t}, \tag{12}$$

where $h(s;\psi)$ is given by (7). The term t^{d-1} comes from the Jacobian of the change of variable from s to $||s||, s/||s||)$.

The upper limit of the integrals in (12) is the largest value of t for which the maximum likelihood estimator corresponding to $s(t)$ exists; in some examples this value can be found analytically, while in general it can be determined numerically.

In the examples of Sections 4·1 and 4·2, φ is the canonical parameter of the exponential family model, and $\psi(\varphi)$ is a nonlinear function of the canonical parameter. In the example of Section 4·3 the assumed model is not an exponential family model, so we use the corresponding tangent exponential model.

In the exponential family examples of Sections 4·1 and 4·2 we compare the directional p-value with Skovgaard's w^* (2), where the general form for γ given in Skovgaard (2001, Eq. (10)) simplifies to his Eq. (13). The latter expression in our notation is

$$\gamma=\frac{\{(s-s_\psi)^\top\jmath^{-1}_{\varphi\varphi}(\widehat{\varphi}_\psi)(s-s_\psi)\}^{d/2}}{w^{d/2-1}(\widehat{\varphi}-\widehat{\varphi}_\psi)^\top(s-s_\psi)}\left\{\frac{|\jmath_{\varphi\varphi}(\widehat{\varphi}_\psi)|}{|\jmath_{\varphi\varphi}(\widehat{\varphi})|}\right\}^{1/2}\left\{\frac{|\jmath_{\lambda\lambda}(\widehat{\varphi}_\psi)|}{|\imath_{\lambda\lambda}(\widehat{\varphi}_\psi)|}\right\}^{1/2}, \tag{13}$$

where $\jmath_{\lambda\lambda}$ and $\imath_{\lambda\lambda}$ are respectively the observed and expected information for the nuisance parameter λ. For calculating the p-value γ is evaluated at $s=0$, corresponding to $y=y^0$.

4. EXAMPLES

4·1. *Comparison of normal means*

Suppose y_{ij} are independent random variables with distributions $N(\mu_i, \sigma_i^2)$, for $i = 1, \ldots, g$, $j = 1, \ldots, n_i$. We want to test the null hypothesis of homogeneity of means among the g groups, i.e.,

$$H_0 : \mu_1 = \cdots = \mu_g,$$

against the alternative that at least one equality does not hold.

The log-likelihood for the parameter $\theta = (\mu_1, \ldots, \mu_g, \sigma_1^2, \ldots, \sigma_g^2)$ is

$$\ell(\theta) = -\frac{1}{2} \sum_{i=1}^{g} \left\{ n_i \log \sigma_i^2 + \frac{1}{\sigma_i^2} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 \right\} . \tag{14}$$

The maximum likelihood estimate is $\widehat{\theta} = (\bar{y}_1, \ldots, \bar{y}_g, v_1^2, \ldots, v_g^2)$, where $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$, $v_i^2 = n_i^{-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$, while the constrained maximum likelihood estimate is $\widehat{\theta}_0 = (\tilde{\mu}, \ldots, \tilde{\mu}, \tilde{v}_1^2, \ldots, \tilde{v}_g^2)$, where $\tilde{v}_i^2 = v_i^2 + (\bar{y}_i - \tilde{\mu})^2$, $i = 1, \ldots, g$, and $\tilde{\mu}$ is the estimate of the common mean which is obtained numerically by maximization of the profile log-likelihood

$$\ell_P(\mu) = \sum_{i=1}^{g} -\frac{n_i}{2} \log\{v_i^2 + (\bar{y}_i - \mu)^2\} .$$

Hence the log-likelihood ratio statistic is

$$w = \sum_{i=1}^{g} n_i \log(\tilde{v}_i^2 / v_i^2),$$

which follows asymptotically the χ^2_{g-1} distribution, under the null hypothesis.

The model (14) is a full exponential family of order $2g$ with canonical parameter $\varphi = (\varphi_1, \ldots, \varphi_{2g})$ and sufficient statistic $s = (u_1, \ldots, u_{2g})$. The components of the canonical parameter are

$$\varphi(\theta)_i = \begin{cases} \mu_i/\sigma_i^2, & i = 1, \ldots, g \\ -1/(2\sigma_i^2), & i = g+1, \ldots 2g \end{cases}$$

while the sufficient statistic has components $u_i = n_i \bar{y}_i$, $u_{g+i} = \sum_{j=1}^{n_i} y_{ij}^2$, for $i = 1, \ldots, g$. The hypothesis of equal means can be written as $H_0 : \psi = 0$, with $\psi_i = \mu_{i+1} - \mu_1$, $i = 1, \ldots, g-1$, with nuisance parameter $\lambda = (\mu, \sigma_1^2, \ldots, \sigma_g^2)$, and $\mu = \mu_1$. This hypothesis places nonlinear constraints on the canonical parameter φ.

For the computation of the directional p-value we need the tilted log-likelihood function (6), $\ell(\varphi; s) = \ell^0(\varphi) + \varphi^\top s$, where $s^0 = 0$ and

$$s_\psi = -\ell^0_\varphi(\widehat{\varphi}_0) = \{n_1(\tilde{\mu} - \bar{y}_1), \ldots, n_g(\tilde{\mu} - \bar{y}_g), 2n_1\tilde{\mu}(\tilde{\mu} - \bar{y}_1), \ldots, 2n_g\tilde{\mu}(\tilde{\mu} - \bar{y}_g)\} .$$

In this example, the log-likelihood along the line $s(t) = ts^0 + (1-t)s_\psi = (1-t)s_\psi$ that joins the expected value s_ψ and the observed value s^0 can be written explicitly in the φ parameteriza-

tion, giving

$$\ell\{\varphi; s(t)\} = \sum_{i=1}^{g} n_i \left[\varphi_i\{\bar{y}_i + (1-t)(\tilde{\mu} - \bar{y}_i)\} + \varphi_{g+i}\{n_i^{-1} \sum_{j=1}^{n_i} y_{ij}^2 + 2(1-t)\tilde{\mu}(\tilde{\mu} - \bar{y}_i)\} + \frac{1}{2}\log(-2\varphi_{g+i}) + \frac{1}{4}\varphi_i^2 \varphi_{g+i}^{-1} \right],$$

which is maximized at

$$\widehat{\varphi}_i(t) = \frac{\tilde{\mu} + t(\bar{y}_i - \tilde{\mu})}{v_i^2 + (1-t^2)(\tilde{\mu} - \bar{y}_i)^2}, \quad \widehat{\varphi}_{g+i}(t) = -\frac{1}{2\{v_i^2 + (1-t^2)(\tilde{\mu} - \bar{y}_i)^2\}}, \quad i = 1, \ldots, g\,.$$

In the θ parameterization we have

$$\widehat{\mu}_i(t) = \tilde{\mu} + t(\bar{y}_i - \tilde{\mu}), \quad \widehat{\sigma}_i^2(t) = v_i^2 + (1-t^2)(\tilde{\mu} - \bar{y}_i)^2 = \tilde{v}_i^2 - t^2(\tilde{\mu} - \bar{y}_i)^2\,, \quad i = 1, \ldots, g\,.$$

As expected, $t = 0$ and $t = 1$ give $\widehat{\theta}_0$ and $\widehat{\theta}$, respectively. Moreover, since $\widehat{\sigma}_i^2(t)$ must be positive for all $i = 1, \ldots, g$, we have that

$$t < t_{\max} = \min_i \sqrt{1 + \frac{v_i^2}{(\tilde{\mu} - \bar{y}_i)^2}};$$

$s(t_{\max})$ is the last value of s along the line $s(t)$ that leads to an admissible maximum likelihood estimate $\widehat{\varphi}\{s(t)\}$.

The directional p-value is computed from (12), with (7) given by

$$h\{s(t); \psi\} \propto \prod_{i=1}^{g} \{\tilde{v}_i^2 - t^2(\tilde{\mu} - \bar{y}_i)^2\}^{(n_i - 3)/2} \left[\sum_{i=1}^{g} \frac{n_i}{(\tilde{v}_i^2)^2} \{\tilde{v}_i^2 - 2t^2(\bar{y}_i - \tilde{\mu})^2\} \right].$$

In the last expression we have used the following results

$$|\jmath_{\varphi\varphi}\{\varphi(\theta); s(t)\}| = \prod_{i=1}^{g} 2n_i^2(\sigma_i^2)^3,$$

$$|\jmath_{\lambda\lambda}\{\varphi(\widehat{\theta}_0); s(t)\}| = \left[\prod_{i=1}^{g} \frac{n_i}{2(\tilde{v}_i^2)^2} \right] \left[\sum_{i=1}^{g} \frac{n_i}{(\tilde{v}_i^2)^2} \{\tilde{v}_i^2 - 2t^2(\bar{y}_i - \tilde{\mu})^2\} \right].$$

Skovgaard (2001)'s modified likelihood ratio statistic w^* can also be computed explicitly for this example, as the correction factor (13) simplifies to

$$\gamma = \left(\prod_{i=1}^{g} \frac{\tilde{v}_i^2}{v_i^2} \right)^{3/2} \frac{\{\sum_{i=1}^{g} n_i(\bar{y}_i - \tilde{\mu})^2/\tilde{v}_i^2\}^{d/2}}{w^{d/2-1}\{\sum_{i=1}^{g} n_i(\bar{y}_i - \tilde{\mu})^2/v_i^2\}} \sqrt{\frac{\sum_{i=1}^{g} n_i(\tilde{v}_i^2)^{-2}\{v_i^2 - (\bar{y}_i - \tilde{\mu})^2\}}{\sum_{i=1}^{g} n_i(\tilde{v}_i^2)^{-1}}}.$$

We illustrate these calculations using the Gravity data given in Example 4.14 of Davison & Hinkley (1997). The data are measures of the acceleration due to gravity, from eight series of experiments, with sample sizes ranging from 8 to 13, and are available in the `R` package `boot`. The series exhibit a clear difference in variability and a possible change in location, as shown in the left panel of Fig. 1. The full and constrained maximum likelihood estimates are given in Table 1. The first order p-value based on the likelihood ratio statistic w is 0.0092. The directional p-value (represented in the right panel of Fig. 1), 0.0336, is much larger, and similar to that obtained with Skovgaard's w^*, which is 0.0320. Both p-values are in agreement with the non-parametric bootstrap p-value 0.030 given in Example 4.14 of Davison & Hinkley (1997). We also

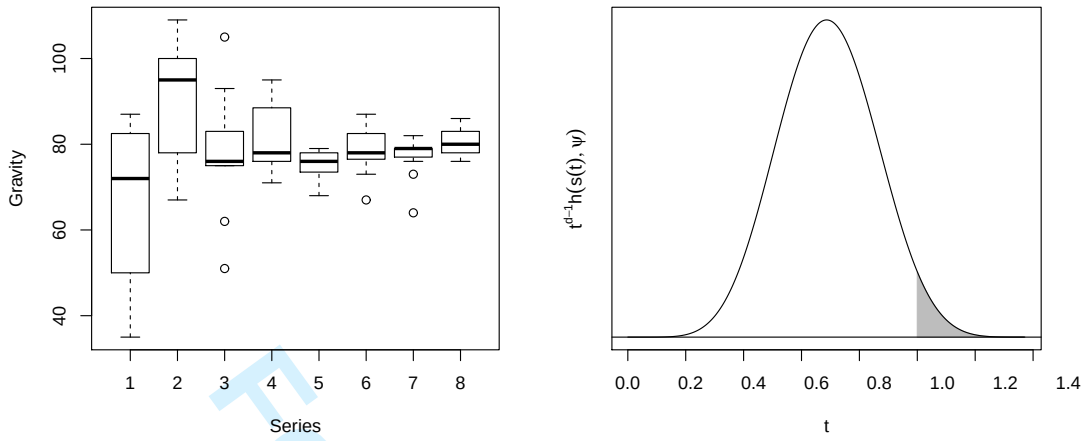

Fig. 1. Gravity data: gravity series box plots (left panel) and integrand function for the directional p-value (12); the gray area represents the directional p-value. (right panel).

Table 1. *Gravity data: Full maximum likelihood estimate (first two rows) and constrained maximum likelihood estimate of the variances under the null hypothesis of common mean (third row), whose estimate is* $\tilde{\mu} = 78.88$.

Series	1	2	3	4	5	6	7	8
$\bar{y}_i$	66.38	89.91	77.33	81.38	75.25	78.91	77.54	80.39
v_i^2	324.23	212.63	220.67	60.23	11.69	30.99	20.71	10.39
$\tilde{v}_i^2$	480.70	334.19	223.07	66.44	24.89	30.99	22.52	12.64

used a parametric bootstrap approximation to the distribution of w, simulating $10,000$ bootstrap samples under the null hypothesis; this gave an estimated p-value of 0.0277.

In this example the directional test and the test based on w^* gave almost identical results, both strongly correcting the p-value from the first-order approximation. However, in more extreme settings this might not be the case. In order to study the differences between the two p-values we performed simulation experiments with balanced samples, varying the number of groups, g, and the number of observations per group, n. These were summarized by comparing the p-values obtained from simulations under the hypothesis to the uniform distribution. For each configuration we considered 10,000 replications, with half of the σ_i^2 equal 1 and half equal to 5, and $\mu = 2$. The results for two cases are reported in Table 2. In the top rows, $n_i = 5$ and $g = 30$, giving 29 parameters of interest and 31 nuisance parameters. The likelihood ratio statistic yields p-values that are too small, but this is corrected by the directional p-value. On the other hand, w^* seems to overcorrect. In the bottom rows we took the more extreme case of $g = 200$ with $n_i = 5$; this has 199 parameters of interest and 201 nuisance parameters. In this case the same pattern is accentuated, but the directional test maintains the level extremely well. This is consistent with the results for linear exponential families in Davison et al. (2014).

Table 2. *Simulated empirical distribution (%) of p-values for testing common means in $g = 30$ groups with $n_i = 5$ observations per group (top rows), and in $g = 200$ groups with $n_i = 5$ observations per group (bottom rows), based on 10,000 replications.*

Nominal (%)	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
Likelihood ratio, (15)	18.5	28.3	38.4	51.3	72.0	89.0	96.6	99.1	99.5	99.8	99.9
Skovgaard's w^*, (2)	0.6	1.4	3.0	7.1	20.1	44.4	71.3	88.4	94.3	96.9	98.7
Directional, (12)	0.9	2.4	4.9	10.0	24.8	50.2	74.9	90.2	95.1	97.5	99.1
Likelihood ratio, (15)	86.5	92.4	95.6	97.9	99.6	99.9	100	100	100	100	100
Skovgaard's w^*, (2)	0.1	0.4	0.9	2.7	10.0	28.7	55.9	78.5	88.1	93.5	96.9
Directional, (12)	1.0	2.5	5.1	10.3	25.2	50.4	74.8	90.0	95.0	97.6	99.1
Standard error	0.1	0.2	0.2	0.3	0.4	0.5	0.4	0.3	0.2	0.2	0.1

4·2. *Marginal independence in Gaussian models*

Let $y_1, \ldots, y_n$ be a sample of independent random vectors from a multivariate normal $N_q(\mu, \Sigma)$, where the mean μ and the covariance matrix Σ are unknown and arbitrary apart from the restriction that Σ is positive definite. In Davison et al. (2014, §5.3) the same model was considered for testing conditional independence hypotheses, i.e. zero entries of the concentration matrix Σ^{-1}, which is linear in the canonical parameter of the exponential family. Here, we use the model to test hypotheses of marginal independence among some variables, i.e. zero entries in Σ, which are nonlinear constraints on the canonical parameter.

Let y denote the $n \times q$ matrix with ith row vector $y_i^\top$. In the following, we consider the likelihood function for Σ, which is based on the Wishart distribution, $W_q(n-1, \Sigma/n)$, of the empirical covariance matrix $y^\top y/n - y^\top 1_n 1_n^\top y/n^2$, and is free of the nuisance parameters μ. This eliminates the nuisance parameter directly, and makes the calculations more transparent, but the same directional test is obtained from the full log-likelihood. (The p-values for the omnibus statistics w and w^* would be slightly less accurate, for instance, w would have n in place of $n-1$.) Then denoting the unbiased estimate of Σ by $S = n(y^\top y/n - y^\top 1_n 1_n^\top y/n^2)/(n-1)$, the log-likelihood function for $\theta = \Sigma$ is

$$\begin{aligned}\ell(\theta) &= -\frac{(n-1)}{2}\log|\Sigma| - \frac{(n-1)}{2}\mathrm{tr}(\Sigma^{-1}S)\,,\\ &= \frac{(n-1)}{2}\log|\varphi(\Sigma)| - \frac{(n-1)}{2}\mathrm{tr}\{\varphi(\Sigma)S\}\,,\end{aligned}$$

where $\varphi = \varphi(\Sigma) = \Sigma^{-1}$ is the canonical parameter of the exponential family. The maximum likelihood estimate is $\widehat{\Sigma} = S$.

Consider now a reduced model in which some off-diagonal elements of Σ equal zero. With ψ denoting the $d \times 1$ vector of these components and λ the remaining $p - d$ parameters, the reduced model corresponds to the null hypothesis $H_0 : \psi = 0$. Under H_0 the constrained maximum likelihood estimate of θ is $\widehat{\theta}_0 = \widehat{\Sigma}_0$, which is typically obtained numerically, for instance using the `R` function `fitCovGraph` in the package `ggm`. The log-likelihood ratio statistic is

$$w = (n-1)\log(|\widehat{\Sigma}_0\widehat{\Sigma}^{-1}|)\,, \tag{15}$$

with limiting χ^2_d distribution as $n \to \infty$.

The value s_ψ from (10) is $s_\psi = -\ell_\varphi\{\varphi(\widehat{\Sigma}_0)\} = (n-1)(\widehat{\Sigma} - \widehat{\Sigma}_0)/2$, and the tilted log-likelihood along the line $s(t) = (1-t)s_\psi$ is

$$\ell\{\theta; s(t)\} = \frac{(n-1)}{2}\log|\varphi(\Sigma)| - \frac{(n-1)}{2}\mathrm{tr}\{\varphi(\Sigma)(t\widehat{\Sigma} + (1-t)\widehat{\Sigma}_0)\},$$

which has maximum $\widehat{\Sigma}(t) = t\widehat{\Sigma} + (1-t)\widehat{\Sigma}_0$. The last value of s along the line $s(t)$, $s(t_{\max})$, is the largest value such that $\widehat{\Sigma}(t)$ is positive definite, and this can easily be found numerically.

The directional p-value (12) uses

$$h\{s(t); \psi\} \propto |t\widehat{\Sigma} + (1-t)\widehat{\Sigma}_0|^{-(n-q-2)/2}\left|\jmath_{\lambda\lambda}\{\widehat{\Sigma}_0; s(t)\}\right|^{1/2},$$

where we have used $\left|\jmath_{\varphi\varphi}\left[\widehat{\varphi}\{s(t)\}; s(t)\right]\right|^{-1/2} = |t\widehat{\Sigma} + (1-t)\widehat{\Sigma}_0|^{-(q+1)/2}$ and where the (j,k)-element of $\jmath_{\lambda\lambda}\{\widehat{\Sigma}_0; s(t)\}$ is given by

$$\jmath_{\lambda_j\lambda_k}\{\widehat{\Sigma}_0; s(t)\} = \frac{(n-1)}{2}\left(\mathrm{tr}(A_{kj}) + t\left[\mathrm{tr}\{(A_{kj} + A_{jk})(\widehat{\Sigma}_0^{-1}\widehat{\Sigma} - I_q)\}\right]\right),$$

with $A_{kj} = \Sigma_0^{-1}(\partial\Sigma/\partial\lambda_k)\Sigma_0^{-1}(\partial\Sigma/\partial\lambda_j)$ and I_q the identity matrix of order q. Since the elements of λ are elements of Σ, $(\partial\Sigma/\partial\lambda_k)$ is a matrix of zeroes with either one or two ones. In particular, if λ_j is a diagonal element of Σ, then $(\partial\Sigma/\partial\lambda_j)$ has one in the corresponding diagonal position; on the other hand, if λ_j is an off-diagonal element of Σ, say of positions (r,s) and (s,r), then $(\partial\Sigma/\partial\lambda_j)$ has one in the corresponding positions.

The correction factor (13) for Skovgaard (2001)'s modified likelihood ratio statistic (2) is

$$\gamma = \frac{\left[\frac{1}{2}\left\{\mathrm{tr}(\widehat{\Sigma}\widehat{\Sigma}_0^{-1}\widehat{\Sigma}\widehat{\Sigma}_0^{-1}) - q\right\}\right]^{d/2}|\widehat{\Sigma}\widehat{\Sigma}_0^{-1}|^{-(q+2)/2}}{\frac{1}{2}\left\{\mathrm{tr}(\widehat{\Sigma}^{-1}\widehat{\Sigma}_0) - q\right\}\left(\log|\widehat{\Sigma}_0\widehat{\Sigma}^{-1}|\right)^{d/2-1}}\left\{\frac{|\jmath_{\lambda\lambda}\{\widehat{\Sigma}_0; s(1)\}|}{|\imath_{\lambda\lambda}(\widehat{\Sigma}_0)|}\right\}^{1/2}, \qquad (16)$$

where the (j,k)-element of $\imath_{\lambda\lambda}(\widehat{\Sigma}_0)$ is $(n-1)\mathrm{tr}(A_{jk})/2$.

As a first illustration we use the data on glucose control of diabetes patients in Cox & Wermuth (1996, page 229). The data were collected on 68 patients with fewer than 25 years of illness to identify psychological and socio-economic variables possibly important for glucose control. The variables are: Y, glucose control, measured by glycosylated haemoglobin; X, knowledge about the illness; Z, U and V measure patients' type of attribution, called fatalistic externality, social externality and internality; and W, the duration of the illness. The data, as well as a more detailed description of the variables, are also available in the `R` package `ggm`. In order to highlight the differences between the various approaches we consider here the subset of 14 observations in $\mathbb{R}^6$ given by males with less than 13 years of formal education and we test mutual independence between the six variables, thus giving ψ of dimension 15 and λ of dimension 6. The first order p-value is 0.0099, the directional p-value is 0.1427, while Skovgaard's w^* gives p-value 0.2632. Mutual independence can also be assessed from the concentration matrix, as in Davison et al. (2014, §5.3). Reassuringly, this yields the same results.

As a second numerical illustration we use the data on gene expression on yeast, analyzed for instance in Chaudhuri et al. (2007, §6.1). There are $n = 134$ experiments with gene expression measurements for eight genes. The empirical covariance matrix is given in the upper triangular part of Table 3. As in Chaudhuri et al. (2007, §6.1) we imposed 13 marginal independence constraints which are indicated by the zeroes in the lower triangular part of Table 3. The first order p-value is 0.0021, the directional p-value is 0.0028, while Skovgaard's w^* gives p-value 0.0026. In this case there are no practical differences between the three approaches. This is largely due to

the relatively large sample size, compared with $q = 8$ variables and the low dimension of both ψ and λ. On the other hand, in contrast to the other examples and to those in Davison et al. (2014), Skovgaard's w^* gives a smaller p-value than the directional approach.

Table 3. *Empirical covariance matrix for the gene expression data on yeast in $n = 134$ experiments (diagonal and upper triangular part) and estimated covariance matrix with marginal independence constraints (lower triangular part). In the diagonal the numbers in parenthesis are the estimates under the reduced model.*

	Y11	Y4	Y80	Y2	Y1	Y3	Y7	Y10
Y11	$^{0.15}_{(0.15)}$	0.03	0.01	-0.12	-0.07	-0.05	-0.05	-0.05
Y4	0.03	$^{0.13}_{(0.13)}$	0.04	-0.02	-0.06	0.03	-0.05	-0.04
Y80	0	0.04	$^{0.22}_{(0.22)}$	0.21	0.22	0.07	0.18	0.19
Y2	0	0	0.06	$^{2.89}_{(2.80)}$	2.51	0.58	2.55	2.28
Y1	0	0	0.09	2.41	$^{2.89}_{(2.76)}$	0.52	2.77	2.41
Y3	0	0	0	0.57	0.49	$^{0.61}_{(0.61)}$	0.72	0.55
Y7	0	0	0	2.49	2.69	0.72	$^{3.42}_{(3.42)}$	2.59
Y10	0	0	0.05	2.20	2.31	0.54	2.54	$^{2.37}_{(2.30)}$

We also investigated the accuracy of the methods by simulation, using the settings in these two examples, but with varying samples sizes. The upper part of Table 4 summarizes a simulation study from the fitted reduced model for the glucose control example. The results underline the high accuracy of the directional approach, while the performances of the first order likelihood ratio test, and Skovgaard (2001)'s adjusted version w^* are respectively poor, and not very accurate. The lower part of Table 4 reports the results when simulating from the fitted model in Table 3, the yeast gene example, for $n = 15$. While with moderate or large sample sizes the accuracy of both directional and Skovgaard's p-values are satisfactory, in extreme settings Skovgaard's w^* indeed tends to give p-values that are too small.

Table 4. *Simulated empirical distribution (%) of p-values for testing hypotheses on the covariance matrix, based on $10,000$ replications. The top rows corresponds to testing mutual independence with $n = 14$ and $q = 6$, as in the glucose control example; the lower rows corresponds to the null model in Table 3, with $n = 15$, as in the yeast gene example; the dimension of the parameter of interest is 15 in the upper rows and 13 in the lower rows.*

Nominal (%)	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
Likelihood ratio, (15)	7.2	13.2	20.3	30.6	51.2	74.4	89.9	96.8	98.6	99.4	99.7
Skovgaard's w^*, (2)	0.8	2.2	4.2	8.5	21.4	43.7	68.7	85.2	91.7	95.5	98.0
Directional, (12)	1.0	2.5	4.9	10.0	25.0	49.1	74.9	89.9	95.0	97.8	99.0
Likelihood ratio, (15)	12.0	19.7	28.0	39.4	60.3	80.9	92.6	97.5	98.9	99.5	99.9
Skovgaard's w^*, (2)	1.4	3.5	6.8	12.6	29.2	54.2	78.0	91.3	95.5	97.7	99.0
Directional, (12)	1.0	2.5	5.4	10.8	26.1	51.2	76.0	90.5	95.2	97.5	99.0
Standard error	0.1	0.2	0.2	0.3	0.4	0.5	0.4	0.3	0.2	0.2	0.1

Finally, we consider a simulation in a rather extreme setting with $q = 50$ variables, in five independent blocks of ten variables each. In each block, the variables are generated from a ten-dimensional equi-correlated normal, with correlation 0.5 and variances ranging from 1 to 5. The null hypothesis assumes independence between the five blocks, while not imposing additional constraints on the other elements of the covariance matrix. The covariance matrix has $1,275$

parameters and the parameter of interest ψ has dimension $1,000$. Table 5 shows the results for different sample sizes ($n = 60, 300, 600$) and they all indicate the usual behaviour with first-order p-values that are too small, directional p-values that are virtually exact and Skovgaard's p-values that are too large.

Table 5. *Simulated empirical distribution (%) of p-values for testing hypotheses on the covariance matrix, based on $10,000$ replications. The null hypothesis assumes independence between five blocks of ten dependent variables each: dimension of the parameter of interest is 1000, with 275 nuisance parameters. The sample size n is 60 (top rows), 300 (middle rows) and 600 (lower rows).*

Nominal (%)	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
Likelihood ratio, (15)	100	100	100	100	100	100	100	100	100	100	100
Skovgaard's w^*, (2)	0.0	0.0	0.0	0.1	0.6	2.6	9.9	22.5	33.3	44.0	57.0
Directional, (12)	0.7	2.4	4.9	9.7	25.0	49.7	74.9	90.4	95.1	97.5	99.0
Likelihood ratio, (15)	25.0	37.9	49.6	63.6	82.2	94.5	98.8	99.8	99.9	100.0	100.0
Skovgaard's w^*, (2)	0.2	0.4	1.0	2.3	8.0	23.2	47.7	70.0	80.7	88.5	94.1
Directional, (12)	1.0	2.5	4.8	9.7	24.8	49.7	74.2	89.9	95.0	97.5	99.1
Likelihood ratio, (15)	6.5	12.2	19.9	31.1	53.8	78.6	92.7	97.9	99.1	99.7	99.9
Skovgaard's w^*, (2)	0.3	1.0	2.2	5.0	15.0	35.2	62.2	81.9	89.4	94.3	97.4
Directional, (12)	1.0	2.4	4.9	9.8	25.1	49.6	75.1	89.8	95.0	97.5	98.9
Standard error	0.1	0.2	0.2	0.3	0.4	0.5	0.4	0.3	0.2	0.2	0.1

4·3. *Box-Cox transformation*

As an example that requires construction of the tangent exponential model, we consider the Box & Cox (1964) model $y_i(\gamma) = x_i^\top \beta + \sigma\, z_i$, $i = 1, \ldots, n$, where

$$y_i(\gamma) = \begin{cases} (y_i^\gamma - 1)/\gamma\,, & \gamma \neq 0, \\ \log y_i\,, & \gamma = 0, \end{cases}$$

x_i is a column vector of dimension $p-2$ of known covariates, $\theta = (\beta, \sigma, \gamma)$ is the unknown parameter and where the error term z_i has a given distribution. For ease of exposition we assume the errors to be standard normal although the method can be extended to non-normal error and to nonlinear regression as well (Fraser et al., 1999).

The log likelihood $\ell(\theta) = \sum_{i=1}^n \ell_i(\theta)$ has

$$\ell_i(\theta) = -\log\sigma - \frac{1}{2\sigma^2}\{y_i(\gamma) - x_i^\top\beta\}^2 + (\gamma - 1)\log y_i\,.$$

The maximum likelihood estimate $\widehat\theta = (\widehat\beta, \widehat\sigma, \widehat\gamma)$ can be found maximizing numerically the profile log likelihood for γ, which is available analytically.

The canonical parameter of the tangent exponential model can be obtained as in Fraser et al. (2009) and is $\varphi(\theta)^\top = \sum_{i=1}^n \{\partial\ell_i(\theta)/\partial y_i\} V_i$, with

$$\frac{\partial\ell_i(\theta)}{\partial y_i} = -\frac{\{y_i(\gamma) - x_i^\top\beta\}}{\sigma^2}\frac{\partial y_i(\gamma)}{\partial y_i} + \frac{\gamma - 1}{y_i}\,,$$

$$V_i = y_i^{1-\widehat\gamma}\left[x_i^\top, \frac{y_i(\widehat\gamma) - x_i^\top\widehat\beta}{\widehat\sigma}, \frac{y_i^{\widehat\gamma} - \widehat\gamma y_i^{\widehat\gamma}\log y_i - 1}{\widehat\gamma^2}\right],$$

where V_i is a row vector of dimension p. Details on these and other quantities used in the following are given in the Appendix.

We consider testing a reduced model in which some components of β, denoted by ψ, have been set to zero. As usual, λ will denote the remaining components of θ. Let $\widehat{\theta}_0$ be the maximum likelihood estimate under the null hypothesis H_0: $\psi = 0$.

The line joining $s^0 = 0$ and s_ψ is $s(t) = (1-t)s_\psi$, with

$$s_\psi = -\left\{\varphi_\theta(\widehat{\theta}_0)^\top\right\}^{-1} \frac{\partial \ell(\theta)}{\partial \theta}\Big|_{\theta=\widehat{\theta}_0} .$$

Hence, the tilted log likelihood along this line is $\ell\{\theta; s(t)\} = \ell(\theta) + \varphi(\theta)^\top s(t)$ and it is maximized numerically, thus giving $\widehat{\theta}\{s(t)\}$ and (minus) the Hessian $\jmath[\widehat{\theta}\{s(t)\}; s(t)]$. These are used to compute

$$|\jmath_{\varphi\varphi}[\widehat{\varphi}\{s(t)\}; s(t)]| = |\jmath[\widehat{\theta}\{s(t)\}; s(t)]|\ |\varphi_\theta[\widehat{\theta}\{s(t)\}]|^{-2} .$$

The only remaining quantity for the computation of (7) is the nuisance parameter adjustment (9)

$$|\jmath_{\lambda\lambda}\{\widehat{\theta}_0; s(t)\}|^{1/2} \propto |\jmath_{\lambda\lambda}(\widehat{\theta}_0) - (1-t)H(\widehat{\theta}_0)| ,$$

where $H(\theta) = \partial\{\varphi_\lambda(\theta)^\top s_\psi\}/\partial\lambda^\top$. While the function within braces is available analytically, we compute its derivative with respect to $\lambda^\top$ numerically, since this has to be done only at $\theta = \widehat{\theta}_0$. Finally, the last value of s along the line $s(t)$, $s(t_{\max})$, is the largest value such that $\widehat{\sigma}(t)$ is positive, and this can easily be found numerically.

As a numerical illustration we consider the data in Table 1 of Box & Cox (1964) which gives the survival times of animals in a 3×4 factorial experiment, where the factors are three poisons and four treatments. Each combination of the two factors is used for four animals, with the allocation to animals being completely randomized. We consider the model with interaction as the full model and we test the null hypothesis of absence of interaction. Thus, the dimension of θ is 14 and the dimension of ψ is 6. The first-order p-value is 0.3219, while the directional p-value is 0.5063, giving even less support to the presence of interaction between the two factors. In this model we have not considered Skovgaard's w^* since its computation is not straightforward.

Figure 2 compares the empirical p-values obtained from simulations under the estimated reduced model to the uniform distribution. As in the previous examples, the directional p-value is extraordinarily accurate, substantially improving upon first-order inference.

5. DISCUSSION

The directional approach to inference for a vector parameter is substantively different than the omnibus approach of the log-likelihood ratio test. In effect, it converts the hypothesized vector parameter value to a scalar parameter, its length, and conditions on the direction of the departure from the tested hypothesis to that indicated by the observed data.

In addition, the use of the saddlepoint approximation underlying the directional approach adjusts for the estimation of nuisance parameters by incorporating the nuisance information function that arises in integrating out the variable that measures the nuisance parameters. This correction is potentially more important in improving the asymptotic properties than the directional version itself. In Example 5.1 in Davison et al. (2014), the directional test gave in simulations results extremely close to Bartlett's test for the homogeneity of variances, which is based on an exact marginal likelihood in which the nuisance parameters in the mean are eliminated. The existence of an exact marginal or conditional likelihood is rare, but the saddlepoint approach which

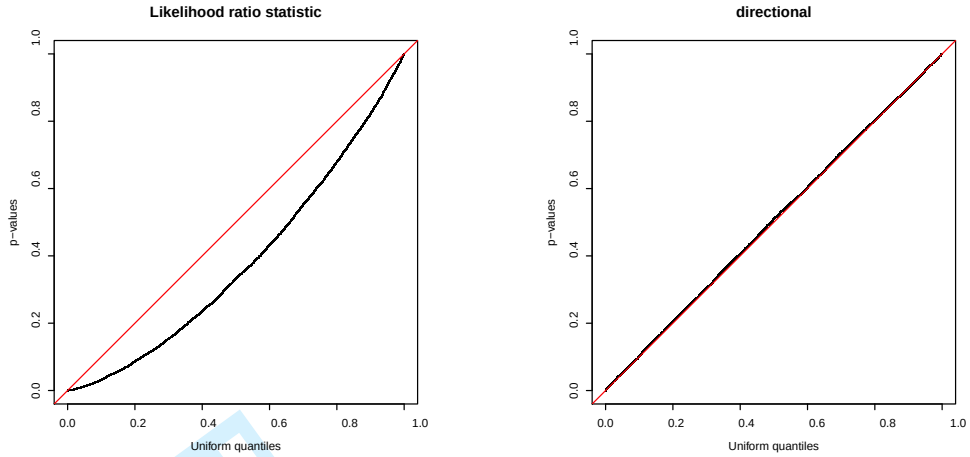

Fig. 2. Box and Cox model. Simulations for testing absence of interaction in a balanced 3×4 factorial model, with 4 replications. We compare the simulated p-values under the null hypothesis, based on 10,000 replications, to the uniform distribution.

is the basis for the tests here seems to capture this adjustment for nuisance parameters, at least approximately.

Finally, as can be seen from the simulations, although the p-value is computed conditionally, it is of course also a marginal p-value, so that the comparison to the omnibus statistics given there is appropriate.

Appendix 1

The tangent exponential model and saddlepoint elimination of nuisance parameters

The exponential family model has the property that its canonical parameter can be obtained by differentiating the log-likelihood function with respect to the sufficient statistic. Sample space derivatives of the log-likelihood function are a key ingredient in higher order asymptotic theory. The tangent exponential model is constructed from the original model by determining a canonical parametrization by the use of ancillary directions. This in effect enable a reduction in dimension of the underlying model, from n to p. These ancillary directions can be represented by a $n \times p$ matrix V, where in the case of independent observations $y_1, \ldots, y_n$, has ith row given by

$$V_i = -\left(\frac{\partial z_i}{\partial y_i}\right)^{-1} \left(\frac{\partial z_i}{\partial \theta^\top}\right)\Big|_{y=y^0, \theta=\widehat{\theta}^0},$$

where $z_i = z_i(y_i; \theta)$ is a pivotal quantity whose distribution is known. One choice of z_i that is always available is $F(y_i; \theta)$, the cumulative distribution function for y_i. In models with an underlying location/scale structure, such as the Box and Cox model, there are usually simpler alternatives. Brazzale et al. (2007, Ch. 8) give examples of the matrix V for a number of common model types.

The matrix V serves to define the canonical parameter of the tangent exponential model:

$$\varphi(\theta)^\top = \sum_{i=1}^{n} \{\partial \ell_i(\theta)/\partial y_i\}\big|_{y^0} \; V_i,$$

and the model itself is

$$f_{\text{TEM}}(s; \varphi) = \exp[\varphi(\theta)^\top s(y) + \ell\{\theta(\varphi); y^0\}]h(s), \tag{A1}$$

where $\ell(\theta; y^0)$ is the observed log-likelihood function for the underlying model for y. The saddlepoint approximation to the density in (A1) is

$$f_{\text{TEM}}(s;\varphi) \doteq \frac{e^{k/n}}{(2\pi)^{p/2}} \exp[\ell(\varphi;s) - \ell\{\widehat{\varphi}(s);s\}]|\jmath_{\varphi\varphi}\{\widehat{\varphi}(s);s\}|^{-1/2}, \tag{A2}$$

where $e^{k/n}$ is a normalizing constant, and $\jmath_{\varphi\varphi}\{\widehat{\varphi}(s);s\}$ is defined at (8).

If the parameter of interest is (a linear combination of) some components of φ, then we can eliminate the remaining components by conditioning on the corresponding sub-vector of s.

If the parameter of interest is a nonlinear function $\psi(\varphi)$, then we construct a density free of the nuisance parameter by finding an approximately ancillary statistic for the nuisance parameter, in the model with ψ fixed (Fraser, 2013; Fraser & Rousseau, 2008; Fraser & Reid, 1995). The conditional density given this ancillary statistic is again of (tangent) exponential form, and thus has a saddlepoint approximation. The ratio of the full model saddlepoint approximation (A2) to this constrained model saddlepoint approximation gives the density of s on $\mathcal{L}_\psi$ as

$$h(s;\psi) \doteq \frac{e^{k'/n}}{(2\pi)^{d/2}} \exp[\ell(\widehat{\varphi}_\psi;s) - \ell\{\widehat{\varphi}(s);s)\}]|\jmath_{\varphi\varphi}\{\widehat{\varphi}(s);s\}|^{-1/2}|\jmath_{(\lambda\lambda)}(\widehat{\varphi}_\psi;s)|^{1/2}, s \in \mathcal{L}_\psi. \tag{3}$$

This result is established by constructing a linear approximation to $\psi(\varphi)$ in order to use linear exponential family theory; this requires a reparameterization as well of the nuisance parameter λ to appropriate complement the parameter of interest. We denote this nuisance reparameterization by (λ) when needed. The final determinant, defined at (9), expresses the information in this re-scaled nuisance parametrization.

The tangent exponential model is defined up to affine transformations of the parameter φ, with a corresponding change in the variable s. If using $h(s;\psi)$ for different values of ψ, rather than a fixed value as considered in this paper, an affine change in φ will have different effects for different values of ψ.

APPENDIX 2

Quantities for the Box and Cox model

We give details on the computation of the needed quantities for the computation of the directional p-value in the Box and Cox model of Section 4·3. The canonical parameter of the tangent exponential model, $\varphi(\theta)^\top = \sum_{i=1}^n \{\partial \ell_i(\theta)/\partial y_i\}\, V_i$, uses the tangent directions, V_i, to the ancillary surface at the observed data point, y^0. Following Fraser et al. (2009), these can be obtained by means of the pivotal quantity $z_i = \{y_i(\gamma) - x_i^\top \beta\}/\sigma$ using the formula

$$V_i = -\left(\frac{\partial z_i}{\partial y_i}\right)^{-1} \left(\frac{\partial z_i}{\partial \theta^\top}\right)\Big|_{y=y^0, \theta=\widehat{\theta}^0},$$

where $\partial z_i/\partial y_i = y_i^{\gamma-1}/\sigma$, $\partial z_i/\partial \beta^\top = -x_i^\top/\sigma$, $\partial z_i/\partial \sigma = -\{y_i(\gamma) - x_i^\top\beta\}/\sigma^2$ and $\partial z_i/\partial\gamma = \{\gamma y_i^\gamma \log y_i - y_i^\gamma + 1\}/(\gamma^2\sigma)$.

The matrix $\varphi_\theta(\theta)$ has generic element $\partial\varphi_r(\theta)/\partial\theta_c$, $r, c = 1, \ldots, p$. In particular,

$$\varphi_\beta(\theta)^\top = \frac{1}{\sigma^2}\sum_{i=1}^n y_i^{\gamma-1} x_i \, V_i \,,$$

$$\varphi_\sigma(\theta)^\top = \frac{2}{\sigma^3}\sum_{i=1}^n y_i^{\gamma-1}\{y_i(\gamma) - x_i^\top\beta\} \, V_i \,,$$

$$\varphi_\gamma(\theta)^\top = -\frac{1}{\sigma^2}\sum_{i=1}^n y_i^{-1}\left[y_i^\gamma \log y_i\{y_i(\gamma) - x_i^\top\beta\} + \gamma^{-2} y_i^\gamma\{\gamma y_i^\gamma \log y_i - y_i^\gamma + 1\} - \sigma^2\right] \, V_i \,,$$

where $\varphi_\beta(\theta)^\top$ is a $(p-2)\times p$ matrix, while $\varphi_\sigma(\theta)^\top$ and $\varphi_\gamma(\theta)^\top$ are p-dimensional row vectors.

Finally, we give the first and second derivatives of $\ell(\theta)$. For notational compactness we will denote by $y(\gamma)$ the column vector with generic element $y_i(\gamma)$ and by X the $n \times p$ matrix with generic row $x_i^\top$. Then we have

$$\begin{aligned}
\frac{\partial\ell(\theta)}{\partial\beta} &= \frac{1}{\sigma^2}X^\top\{y(\gamma) - X\beta\} \,, \\
\frac{\partial\ell(\theta)}{\partial\sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\{y(\gamma) - X\beta\}^\top\{y(\gamma) - X\beta\} \,, \\
\frac{\partial\ell(\theta)}{\partial\gamma} &= -\frac{1}{\gamma^2\sigma^2}\sum_{i=1}^n\{y_i(\gamma) - x_i^\top\beta\}\{\gamma y_i^\gamma \log y_i - y_i^\gamma + 1\} + \sum_{i=1}^n \log y_i \,, \\
\frac{\partial^2\ell(\theta)}{\partial\beta\partial\beta^\top} &= -\frac{1}{\sigma^2}X^\top X \,, \\
\frac{\partial^2\ell(\theta)}{\partial\beta\partial\sigma} &= -\frac{1}{\sigma^3}X^\top\{y(\gamma) - X\beta\} \,, \\
\frac{\partial^2\ell(\theta)}{\partial\beta\partial\gamma} &= \frac{1}{\gamma^2\sigma^2}\sum_{i=1}^n\{\gamma y_i^\gamma \log y_i - y_i^\gamma + 1\} \, x_i \,, \\
\frac{\partial^2\ell(\theta)}{\partial(\sigma)^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\{y(\gamma) - X\beta\}^\top\{y(\gamma) - X\beta\} \,, \\
\frac{\partial^2\ell(\theta)}{\partial\sigma\partial\gamma} &= \frac{2}{\gamma^3\sigma^2}\sum_{i=1}^n\{y_i(\gamma) - x_i^\top\beta\}\{\gamma y_i^\gamma \log y_i - y_i^\gamma + 1\} \,, \\
\frac{\partial^2\ell(\theta)}{\partial(\gamma)^2} &= -\frac{1}{\gamma^2\sigma^4}\sum_{i=1}^n\{\gamma y_i^\gamma \log y_i - y_i^\gamma + 1\}^2 \\
&\quad + \{y_i(\gamma) - x_i^\top\beta\}\{\gamma^3 y_i^\gamma (\log y_i)^2 - 2\gamma^2 y_i^\gamma \log y_i + 2\gamma y_i^\gamma - 2\gamma\} \,.
\end{aligned}$$

References

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics*. London: Chapman & Hall.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. London A* **160**, 268–282.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* **26**, 211–252.

Brazzale, A. R., Davison, A. C. & Reid, N. (2007). *Applied Asymptotics*. Cambridge: Cambridge University Press.

Chaudhuri, S., Drton, M. & Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika* **94**, 199–216.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.

Cox, D. R. & Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. London: Chapman & Hall.

DAVISON, A. C., FRASER, D. A. S., REID, N. & SARTORI, N. (2014). Accurate directional inference for vector parameters in linear exponential families. *Journal of the American Statistical Association* **109**, 302–314.
DAVISON, A. C. & HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
FRASER, D. (2013). Why does statistics have two theories? In *Past, Present and Future of Statistical Science*, X. Lin, C. Genest, D. Banks, G. Molenberghs, D. Scott & J.-L. Wang, eds. Boca Raton: Chapman & Hall, pp. 237–252.
FRASER, D. A. S. & MASSAM, H. (1985). Conical tests: Observed levels of significance and confidence regions. *Statist. Hefte* **26**, 1–17.
FRASER, D. A. S. & REID, N. (1995). Ancillaries and third order significance. *Utilitas Mathematica* **47**, 33–53.
FRASER, D. A. S. & ROUSSEAU, J. (2008). Studentization and deriving accurate p-values. *Biometrika* **95**, 1–16.
FRASER, D. A. S., WONG, A. & SUN, Y. (2009). Three enigmatic examples and inference from likelihood. *Canadian Journal of Statistics* **37**, 161–181.
FRASER, D. A. S., WONG, A. & WU, J. (1999). Regression Analysis, Nonlinear or Nonnormal: Simple and Accurate p Values from Likelihood Analysis. *Journal of the American Statistical Association* **94**, 1286–1294.
SKOVGAARD, I. M. (1988). Saddlepoint expansions for directional test probabilities. *J. R. Statist. Soc. B* **50**, 3–32.
SKOVGAARD, I. M. (2001). Likelihood asymptotics. *Scand. J. Statist.* **28**, 3–32.