Accurate directional inference for vector parameters

A. C. Davison, D. A. S. Fraser, N. Reid and N. Sartori*

February 7, 2013

Abstract

We consider a vector-valued parameter of interest in the presence of a finite-dimensional nuisance parameter, based on higher order asymptotic theory for likelihood inference. We propose a directional test for the vector parameter of interest, that is computed using one-dimensional integration. For discrete responses this extends the development of Davison et al. (2006), and several examples below concern testing hypotheses in contingency tables. For continuous responses the work extends the directional test of Cheah et al. (1994). Exponential family examples and simulations illustrate the high accuracy of the method, which we compare with an adjusted likelihood ratio test of Skovgaard (2001). In a high-dimensional covariance selection example the approach works essentially perfectly, whereas its competitors fail catastrophically.

Keywords: Bartlett Factor; Components of Variance; Contingency Table; Covariance Selection; Exponential Family Model; Graphical Models; Higher-order Asymptotics; Likelihood Ratio Test.

*Anthony Davison is Professor of Statistics, EPFL-FSB-MATHAA-STAT, Station 8, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland, `Anthony.Davison@epfl.ch`. Nancy Reid and Don Fraser are Professors of Statistics, Department of Statistics, University of Toronto, Toronto, Canada M5S 3G3, `dfraser@utstat.toronto.edu`, `reid@utstat.toronto.edu`. Nicola Sartori is Professor of Statistics, Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Via Cesare Battisti 241, 35121 Padova, Italy, `Nicola.Sartori@unipd.it`. This research was partially supported by the Swiss National Science Foundation, the Canadian Natural Sciences and Engineering Research Council, the Senior Scholars funding from York University, Canada, and the Cariparo Foundation Excellence Grant.

1 Introduction

The likelihood ratio statistic is probably the most widely used approach to the comparison of nested parametric models—for example, deviance tests in generalized linear models (McCullagh and Nelder, 1989) are of this type—and provides a general and powerful framework for such comparisons. In large samples chi-squared approximations to the distribution of the likelihood ratio statistic may be used, but in certain cases, for example in sparse contingency tables or covariance selection models, the accuracy of these approximations may be poor. Thus it is of wide interest to consider alternative potentially more accurate approaches.

In this paper we discuss a directional approach derived from higher order approximations for likelihood inference. For a scalar parameter of interest a pivotal quantity, often called r^*, can be constructed, which follows a standard normal distribution with relative error $O(n^{-3/2})$, when the response y is continuous, and with relative error $O(n^{-1})$, when y is discrete. Since these approximations have bounded relative error both in the centre of the distribution and in large deviation regions, they provide highly accurate inferences well into the distribution tails. A review of this literature and several examples are given in Brazzale et al. (2007) and Brazzale and Davison (2008); the discrete case is considered in more generality in Davison et al. (2006).

A development for vector parameters of interest, parallel to that of r^*, was given in Skovgaard (2001). The resulting test statistic has a distribution close to χ^2 and was derived analogously to r^*, so that the approximation is also accurate in large deviation regions. The present paper provides an alternative highly accurate approximation that improves on the usual likelihood ratio statistic, seems to be more accurate in simulations than Skovgaard's statistic, and is very easy to compute.

Our approach starts with a vector-valued measure of departure from the hypothesis, and computes p-values based on the magnitude of this measure, conditional on its direction, thus generalizing one-sided tests for a scalar parameter of interest. Directional tests for vector parameters of interest were proposed in Fraser and Massam (1985) and Skovgaard (1988). For exponential family models, the sufficient statistic for the parameter provides the starting point for this vector measure, as proposed by Cheah et al. (1994), and the development is based on the saddlepoint approximation

to the conditional distribution. The p-value is computed by one-dimensional numerical integration, evaluated conditionally on the direction of the variable; see (10) and (11) below. In this paper we consider exponential family models where the parameter of interest is linear in the canonical parameter. Not only does this encompass many important models, but other approximations are available with which our approach can be compared, thus giving a broad indication of its likely quality when extended to more general settings. The examples below include multi-dimensional contingency tables, binary regression, comparison of variances in normal models and rate parameters in exponential models, and inference about the concentration matrix in graphical models. In simulations the proposed approach is shown to be very accurate, even when inference on the likelihood ratio statistic fails. The method also captures the structure of the models, for example reproducing the F-test for comparing two means in Example 5.3.

2 Background

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

We write $\psi(\theta)$ for the d-dimensional parameter of interest, and consider inference for ψ by assessing the hypothesis $H_\psi : \psi(\theta) = \psi$. In several examples $\theta = (\psi, \lambda)$, i.e., ψ is a component of the full parameter, possibly after re-parameterization. We let $\hat{\theta}_\psi$ denote the constrained maximum likelihood estimator of θ under H_ψ; in component form $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$.

To a first order of approximation, $\hat{\theta}$ follows a normal distribution with mean θ and covariance matrix estimated by $\jmath^{-1}(\hat{\theta})$, where $\jmath(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^\top$ is the observed Fisher information function; an analogous result holds for $\hat{\theta}_\psi$ under H_ψ (Cox and Hinkley, 1974, Ch. 9.3). A parameterization-invariant measure of departure of $\hat{\theta}$ from H_ψ is given by the likelihood ratio

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

Table 1: Retarded activity amongst psychiatric patients (Everitt, 1992, Table 3.3)

	Affective disorders	Schizophrenics	Neurotics
Retarded	12	13	5
Not retarded	18	17	25

With relative error $O(n^{-1})$, $w(\psi)$ follows a χ^2_d distribution, with degrees of freedom d equal to the number of constrained parameters in H_ψ. The apparent improvement from $O(n^{-1/2})$ for the distribution of the maximum likelihood estimator to $O(n^{-1})$ for the likelihood ratio statistic is somewhat artificial; if $d = 1$ the $O(n^{-1/2})$ terms in the error in each tail of the distribution cancel, but one-sided inferences do not improve.

Skovgaard (2001) attributes the exceptional accuracy of the r^* approximation for inference about a scalar interest parameter both to the relative error in the approximation and to its large deviation properties, and proposes an analogous version for vector interest parameters designed to maintain accuracy in the tails of the distribution. The resulting quantity

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

uses a correction factor γ that compares w to an asymptotically equivalent quadratic form. Skovgaard (2001) shows that in addition to having good large-deviation properties, $w^*(\psi)$ is also easier to calculate than the Bartlett adjustment discussed in §6.

Like the likelihood ratio $w(\psi)$, (2) gives an omnibus measure of departure; all potential directions away from the hypothesis H_ψ are averaged in the calculation of p-values. We propose a measure of departure that incorporates information in the data about the relevant direction of deviation from H_ψ, by conditioning. Some comparison of omnibus and directional tests is given in Fraser and Reid (2006).

We consider testing independence for the data in Table 1 to illustrate the ideas in a context in which they can readily be visualized. The nuisance parameter, λ is four-dimensional, consisting of the intercept, one row effect and two parameters for column effects, which are eliminated from inference by conditioning on the table margins. The full model has an additional two-dimensional parameter of interest, ψ, representing the

interaction between rows and columns, and the hypothesis of independence is $H_0 : \psi = 0$. Both models can easily be fitted using software for generalized linear models.

We measure departure from H_ψ on a line in the sample space, indexed by $t \in \mathbb{R}$. As t varies from zero to its maximum possible value, the magnitude of departure varies from the null hypothesis, through the observed table, and through other 2×3 tables with the same margins. Four of these tables are indicated in the right-hand side of Figure 1: the independence table, $t = 0$, an intermediate table, $t = 0.5$, the observed table, $t = 1$, and the most extreme table consistent with the margins, $t = 2$. The upper left panel shows the density $h(t)$, given in (10) below, on this line, with points $t = 0, 0.5, 1, 2$ indicated. The lower left panel shows the shape of the relative density $t^{d-1}h(t)$, for $t > 0$, used in (11) to compute the directional p-value.

The directional p-value (11) is computed using one-dimensional numerical integration and equals 0.050; the first-order p-value obtained using the asymptotic χ^2_2 distribution of the likelihood ratio statistic is 0.047. Skovgaard (2001)'s w^* gives 0.048, and a conditional simulation using the method of Kolassa and Tanner (1994) gives 0.051. The sample size in this example is too large for the methods to give very different p-values.

In the next section we give the details for the directional approximation, and in §4 illustrate its accuracy on some larger contingency tables.

3 Directional tests in linear exponential families

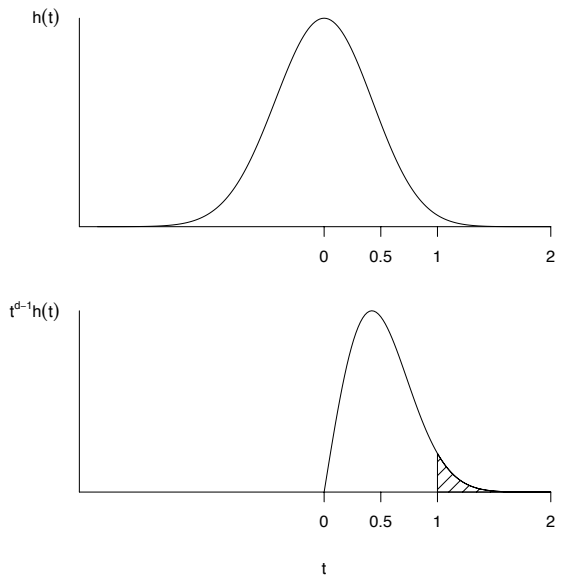
We assume that the model is an exponential family with canonical parameter $\varphi = \varphi(\theta)$ and score variable, or sufficient statistic, $u = u(y)$,

$$f(y;\theta) = \exp[\varphi(\theta)^\top u(y) - K\{\varphi(\theta)\}]h(y). \tag{3}$$

Since u is sufficient for φ, the log-likelihood function is equivalently obtained from the marginal density for u, and may be written as

$$\ell(\theta;u) = \varphi(\theta)^\top u - K\{\varphi(\theta)\} = \varphi(\theta)^\top(u - u^0) + \log f(u^0;\theta), \tag{4}$$

where y^0 is the observed value of the data, with maximum likelihood estimate $\hat\theta^0$, and observed sufficient statistic $u^0 = u(y^0)$. It is convenient in what follows to use centered



$t = 0$

10	10	10
20	20	20

$t = 0.5$

11.0	11.5	7.5
19.0	18.5	22.5

$t = 1$

12	13	5
18	17	25

$t = 2$

14	16	0
16	14	30

Figure 1: Directional inference for the data in Table 1. The conditional density $h(t)$ on the line indexed by t (top left) and the directed radial distance from the value expected under independence, $t^{d-1}h(t)$ (bottom left). The shaded area represents the directional p-value. On the right side the expected data under the hypothesis ($t = 0$) and the observed data ($t = 1$) are indicated, together with the expected data for an intermediate case ($t = 0.5$) and the boundary case ($t = t_{\max} = 2$).

sufficient statistics, so we let $s = u - u^0$, and write

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

where $\ell^0(\theta) = \ell(\theta; s = 0) = \ell(\theta; u = u^0)$. The function $\ell(\theta; s)$, which we call the tilted log-likelihood function, is the key ingredient for the calculation of directional p-values, and the centering means that the observed value of s is $s^0 = 0$.

We further assume that both the parameter of interest, and the nuisance parameter are linear in φ, so $\varphi = \theta = (\psi, \lambda)$, and thus write

$$\ell(\varphi; s) = \psi^\top s_1 + \lambda^\top s_2 + \ell^0(\psi, \lambda), \tag{6}$$

where ψ and s_1 are d-dimensional. The constrained maximum likelihood estimate under H_ψ is $\hat{\varphi}_\psi = (\psi, \hat{\lambda}_\psi)$, and when computed at y^0 it is denoted by $\hat{\varphi}^0_\psi$. Derivatives of $\ell(\cdot)$ are shown by subscripts; for example, $\ell_\varphi(\varphi; s) = \partial\ell(\varphi; s)/\partial\varphi$. Inference for ψ is based on the conditional density for s_1 given s_2, and the saddlepoint approximation to this can be expressed as

$$\hat{f}(s; \psi)\,\mathrm{d}s = c \exp[\ell(\hat{\varphi}^0_\psi; s) - \ell\{\hat{\varphi}(s); s\}] |\jmath_{\varphi\varphi}\{\hat{\varphi}(s); s\}|^{-1/2}\,\mathrm{d}s, \quad s \in \mathcal{L}^0, \tag{7}$$

where c is a constant and $\mathcal{L}^0$ is a d-dimensional plane defined by setting $s_2 = 0$, or equivalently setting $\hat{\lambda}_\psi = \hat{\lambda}^0_\psi$. In (7) $\hat{\varphi}(s)$ is obtained from (5) or (6) as the solution of $s = -\ell^0_\varphi\{\hat{\varphi}(s)\}$, and $\jmath_{\varphi\varphi}(\varphi; s) = -\partial^2\ell(\varphi; s)/\partial\varphi\partial\varphi^\top$. Although it is more conventional to write $\hat{f}(s_1 \mid s_2; \psi)$ or a similar expression for the saddlepoint approximation to the conditional density, the conditioning is here implicitly accommodated by taking a 'slice' through the full density, i.e. constraining s to lie in $\mathcal{L}^0$. The saddlepoint approximation to the conditional density is derived in Barndorff-Nielsen and Cox (1979) and for generalized linear models discussed in Davison (1988). A direct derivation and presentation entirely in terms of likelihood is given in Fraser (2012), and generalized there to inference for nonlinear functions of the canonical parameters, and to approximate exponential models.

We now define a line $\mathcal{L}^*$, in $\mathcal{L}^0$, joining the observed value of s, which is $s^0 = 0$, and its expected value s_ψ under H_ψ; from (6)

$$s_\psi = -\ell^0_\varphi(\hat{\varphi}^0_\psi) = \begin{bmatrix} -\ell^0_\psi(\hat{\varphi}^0_\psi) \\ 0 \end{bmatrix}; \tag{8}$$

note that the value of s_ψ depends on y^0. We parameterize this line by $t \in \mathbb{R}$,

$$s(t) = s_\psi + t(s^0 - s_\psi) = (1-t)s_\psi; \tag{9}$$

the maximum likelihood estimates $\hat{\phi}(s)$ in (7) vary with $s(t)$. As t increases they trace out a curve in the parameter space that passes through the constrained maximum likelihood estimate $\hat{\varphi}^0_\psi$ when $t = 0$ and through the full maximum likelihood estimate $\hat{\varphi}^0$ when $t = 1$.

The conditional density approximation (7) constrained to $\mathcal{L}^*$ is simply

$$h(t;\psi) = \hat{f}\{s(t);\psi\} = c \exp\Big(\ell\{\hat{\varphi}^0_\psi; s(t)\} - \ell[\hat{\varphi}\{s(t)\}; s(t)]\Big) \big| j_{\varphi\varphi}\left[\hat{\varphi}\{s(t)\}; s(t)\right] \big|^{-1/2}. \tag{10}$$

This expression does not require an explicit parametrization of the nuisance parameter for its computation, if we use the more general form $\hat{\varphi}^0_\psi = \arg\sup_{\psi(\varphi)=\psi} \ell^0(\varphi)$ to define the constrained maximum likelihood estimator. This is useful for the examples considered in §5.

Transforming the density for s on the plane $\mathcal{L}^0$ to the conditional density of $\|s\|$, given $s/\|s\|$, is a change from vector to spherical coordinates, so introduces a Jacobian proportional to t^{d-1} (Fraser and Massam, 1985); the relevant dimension is d because $s_2 = 0$ on $\mathcal{L}^0$; see the Supplementary Notes. The directional test computes the p-value as the probability, from $h(t;\psi)$, that $s(t)$ is as far or farther from s_ψ than is the observed value 0; this distribution is on the part of $\mathcal{L}^*$ for which $t > 0$. The directed p-value is thus

$$p(\psi) = \frac{\int_1^{t_{\max}} t^{d-1} h(t;\psi)\,\mathrm{d}t}{\int_0^{t_{\max}} t^{d-1} h(t;\psi)\,\mathrm{d}t}, \tag{11}$$

where $t = 0$ and $t = 1$ correspond respectively to $s = s_\psi$ and to the observed value $s^0 = 0$. This is a refinement of the approach that uses $2\min\{p(\psi), 1-p(\psi)\}$ in the scalar parameter case, described for example in Cox and Hinkley (1974, Ch. 3). The density $h(t;\psi)$ and the function $t^{d-1}h(t;\psi)$ are illustrated in Figure 1.

The upper limit of the integrals in (11) is the largest value of t for which the maximum likelihood estimator corresponding to $s(t)$ exists; for instance, $t_{\max} = 2$ in the example of Figure 1, though $t_{\max}$ may be infinite in some cases. Figure 2 shows the log-likelihood function $\ell[\hat{\varphi}\{s(t)\}; s(t)]$ at four different values of t, including the

observed table, $t = 0$, an intermediate case $t = 0.5$, the value under the hypothesis of independence, $t = 1$, and the extreme case $t = t_{\max}$. These log-likelihood functions correspond to the four 3×2 tables shown in the right column of Figure 1.

The theoretical accuracy of the approximation (11) stems from that of the renormalized saddlepoint approximation (7), so there is at worst a relative error of $O(n^{-1})$ (Butler, 2007, p. 112), even in large deviation regions, and in local deviation regions, which are of most statistical interest, the relative error is $O(n^{-3/2})$ for continuous responses. In some cases the accuracy may even be better, perhaps because of the ratio of similar integrals in (11); for example, in the normal distribution settings of §5.1, the approximation seems to be essentially exact.

4 Models with discrete responses

4.1 Contingency tables

The calculations are particularly straightforward for a generic contingency table, as in the example in §2. Denote the observed cell frequencies by $y^0 = (y_1, \ldots, y_C)$, where C is the total number of cells in the table; for instance, $C = IJ$ in a two-way contingency table with I rows and J columns. With X and θ denoting the $C \times p$ design matrix and the $p \times 1$ parameter vector, we assume a log linear model for the cell frequencies with expected value $\mu(\theta) = \exp(X\theta)$.

The model is a linear exponential family with canonical parameter $\varphi = \theta$ and observed log-likelihood function

$$\ell^0(\varphi) = \varphi^\top X^\top y^0 - 1_C^\top e^{X\varphi}\,, \tag{12}$$

where 1_C is a $C \times 1$ vector of ones and $X^\top y$ is the minimal sufficient statistic. The score function and the observed information are respectively

$$\begin{aligned} \ell^0_\varphi(\varphi) &= X^\top(y^0 - e^{X\varphi}) = X^\top\{y^0 - \mu(\varphi)\}, \\ \jmath_{\varphi\varphi}(\varphi) &= X^\top \mathrm{diag}\{e^{X\varphi}\}X = X^\top \mathrm{diag}\{\mu(\varphi)\}X\,. \end{aligned} \tag{13}$$

For inference about a component parameter ψ of φ, the columns of the design matrix are partitioned as $X = [X_1 \;\; X_2]$, in conformity with $\varphi = (\psi, \lambda)$. The hypothesis

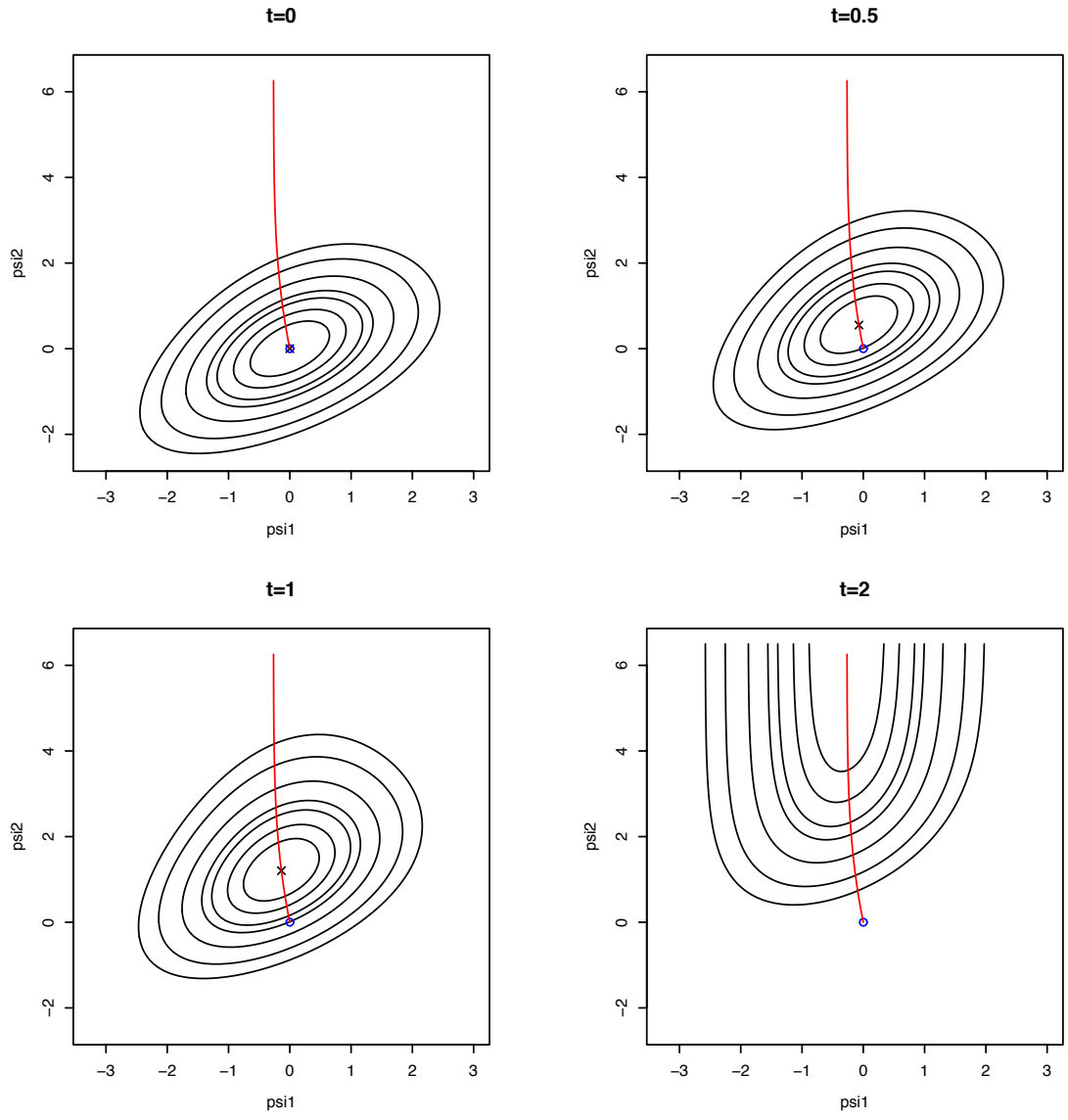


Figure 2: The tilted log-likelihood function, (6), used in (11), and the curve traced by the maximum likelihood estimate ($\times$) as t increases from $t = 0$ ($\circ$) and $s(t)$ varies along $\mathcal{L}^*$ for the case $d = 2$, as in Table 1. Each figure corresponds to a 2×3 table indicated in the right column of Figure 1.

$H_0 : \psi = 0$ corresponds to the equivalence of the two nested models with linear predictors $X_2\lambda$ and $X\varphi$. The constrained maximum likelihood estimate of φ satisfies $\ell_\lambda(\hat{\varphi}_\psi; y^0) = X_2^\top(y^0 - e^{X\hat{\varphi}_\psi}) = 0$.

For directional assessment of the null hypothesis, the observed data point $s^0 = 0$ and the expected value s_ψ defined in (8) is

$$s_\psi = \begin{bmatrix} -X_1^\top(y^0 - e^{X\hat{\varphi}_\psi}) \\ 0 \end{bmatrix} = -X^\top\{y^0 - \mu(\hat{\varphi}_\psi)\}\,.$$

The directional p-value is obtained numerically from (11). To determine $t_{\max}$, note that the maximum likelihood estimate satisfies the condition

$$X^\top y^0 = X^\top \mu(\hat{\varphi}^0)\ : \tag{14}$$

the observed value of the sufficient statistic equals the expected value under the assumed model. In a contingency table, (14) implies that some marginal totals are equal in the observed table and in the fitted table (Birch, 1963). Moreover, if some of these totals are zero, then the maximum likelihood estimate will lie on the boundary of the parameter space (see, e.g., Agresti, 2002, §9.8.2). When we need to compute $\hat{\varphi}(t)$, which maximizes $\ell(\varphi; t) = \ell^0(\varphi) + \varphi^\top s(t)$, equation (14) becomes

$$X^\top\{\hat{\mu}_\psi^0 + t(\hat{\mu}^0 - \hat{\mu}_\psi^0)\} = X^\top\hat{\mu}(t)\,, \tag{15}$$

where $\hat{\mu}^0 = \mu(\hat{\varphi}^0)$, $\hat{\mu}_\psi^0 = \mu(\hat{\varphi}_\psi^0)$ and $\hat{\mu}(t) = \mu\{\hat{\varphi}(t)\}$. For any given value of t larger than 1, the maximum likelihood estimate $\hat{\varphi}(t)$ and corresponding mean parameter $\hat{\mu}(t)$ can be easily obtained by solving (15) using iteratively reweighted least squares. A value of t will be admissible if the corresponding fitted cell frequencies $\hat{\mu}(t)$ are all non-negative and the marginal totals implied by (15) are all positive; $t_{\max}$ is the largest such value of t.

Furthermore, if the larger model is saturated, X will be an invertible matrix of dimension $C \times C$. Then (15) simplifies to

$$\hat{\mu}(t) = \hat{\mu}_\psi^0 + t(\hat{\mu}^0 - \hat{\mu}_\psi^0)\,,$$

and for the value of t to be admissible each element of $\hat{\mu}(t)$ must be positive; i.e.

$$t < t_{\max} = \min_{i;(\hat{\mu}_\psi^0 - \hat{\mu}^0)_i > 0} \frac{(\hat{\mu}_\psi^0)_i}{(\hat{\mu}_\psi^0)_i - (\hat{\mu}^0)_i}.$$

Table 2: Sexual enjoyment data (Kolassa and Tanner, 1994, §3.1)

	Wife's response			
Husband's response	Never or occasionally	Fairly often	Very often	Almost always
Never or occasionally	7	7	2	3
Fairly often	2	8	3	7
Very often	1	5	4	9
Almost always	2	8	9	14

The directional p-value is obtained from (11) with $h(t;\psi)$ given in (10), here equal to

$$h(t;\psi) = \exp\left[1_C^\top\{\hat\mu(t) - \hat\mu_\psi^0\} - \hat\mu(t)^\top\{\log\hat\mu(t) - \log\hat\mu_\psi^0\}\right] |X^\top \mathrm{diag}\{\hat\mu(t)\}X|^{-1/2}\,.$$

As a second example we use the data in Table 2. The structure of the model is similar to that for the data in Table 1, but here λ and ψ have dimensions 7 and 9; again we test the null hypothesis $H_0 : \psi = 0$ of independence. The directional p-value (11) is 0.139, while the first-order and Skovgaard's w^* p-values are respectively 0.078 and 0.165. Kolassa and Tanner (1994, §3.1) reported a simulated conditional p-value of 0.111.

Finally we consider the data in Table 9.17 of Agresti (2002, p. 401), which describes the joint distribution of four dichotomous variables: age of mother (A), length of gestation (G), infant survival (I) and number of cigarettes smoked per day during gestation (S). It is appropriate to treat length of gestation and infant survival as responses and the other variables as explanatory. As a null model we take that with all main effects and three first-order interactions (IG, IA and SA); this has an 8-dimensional parameter λ consisting of the intercept, all four main effects and three first-order interactions. A larger model includes two additional first-order interaction parameters, IS and GA. The directional p-value (11) for testing equivalence of the two models is 0.050, while the first-order p-value based on a chi-squared approximation is 0.052, and Skovgaard's w^* gives p-value 0.048.

4.2 Simulations

In each example above, the p-value from the χ^2 approximation to the likelihood ratio statistic is slightly smaller than the directional p-value, although not enough to make a practical difference. We performed some simulations to investigate the accuracy of the directional p-values, when examined unconditionally. The first set of simulations was based on Table 1: 100,000 2×3 tables, with total sample size $n = 90$, were generated from the independence model. Table 3 shows that the directional p-values are unconditionally very accurate, as are Skovgaard (2001)'s large deviation version w^* and Pearson's χ^2 statistic. The likelihood ratio statistic has the worst performance in this setting.

We increased the parameter dimensions by simulating 4×4 tables; there are seven nuisance parameters and nine interest parameters which when equal to zero yield independence of the row and column classifications. In 100,000 simulations with total sample size $n = 150$, 4,747 of the simulated tables had a cell margin of zero, in which case neither the directional method nor Skovgaard (2001)'s method can be used. In these cases we substituted the first-order likelihood ratio test when computing the simulated p-values. Again both w^* and the directional test give very accurate results, improving on both Pearson's χ^2 and the likelihood ratio test.

The final simulation tests independence in a $6 \times 3 \times 2$ table, with total sample size $n = 1000$. Such a large sample size is needed to avoid too many simulations with zeros in the margins; $14,417$ of 100,000 simulated tables had at least one marginal zero. In such cases the simulation p-values were again computed using the χ^2_1 approximation to the likelihood ratio statistic $w(\psi)$. In this setting there are 27 parameters of interest, with nine nuisance parameters. The directional test and Skovgaard (2001)'s large deviation test again largely retain their accuracy, though the large number of cases in which $w(\psi)$ must be used leads to some deterioration in the lower tail.

The differences between the approximations are small in all three cases, and here Skovgaard (2001)'s large deviation statistic and the directional test yield essentially identical p-values. This is not the case in general, however, as is seen in §5.

Table 3: Comparison of p-values (%) for tests of independence in $100,000$ simulated contingency tables of dimensions 2×3, 4×4 and $6 \times 3 \times 2$. For the 4×4 tables, the likelihood ratio statistic was used instead of (11) or (2) for $4,747$ tables with zero counts in the margins. For the $6 \times 3 \times 2$ tables replacement took place for $14,417$ such tables.

Dimension	Nominal	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
2×3	Lik. Ratio, (1)	1.1	2.8	5.5	10.7	26.0	51.0	75.7	90.5	95.2	97.4	99.2
	Pearson's χ^2	0.9	2.4	5.1	10.3	25.7	50.8	75.7	90.5	95.2	97.4	99.2
	Skovgaard's $w^*(\psi)$, (2)	1.0	2.5	5.0	10.1	25.1	50.1	75.2	90.0	95.0	97.3	99.1
	Directional, (11)	1.0	2.4	5.0	10.0	25.0	50.1	75.2	90.2	95.0	97.3	99.2
4×4	Lik. Ratio, (1)	1.4	3.4	6.4	12.3	28.5	53.6	77.3	91.1	95.5	97.7	99.1
	Pearson's χ^2	0.9	2.3	4.8	9.9	25.5	51.2	76.2	90.8	95.4	97.7	99.1
	Skovgaard's $w^*(\psi)$, (2)	1.1	2.7	5.2	10.2	25.1	49.7	74.6	89.8	94.8	97.3	98.9
	Directional, (11)	1.1	2.6	5.1	10.0	24.8	49.5	74.6	89.9	94.9	97.4	98.9
$6 \times 3 \times 2$	Lik. Ratio, (1)	1.5	3.6	6.9	12.9	29.5	54.8	78.3	91.6	95.9	98.0	99.2
	Pearson's χ^2	1.0	2.5	5.1	10.4	26.1	52.1	77.0	91.1	95.7	97.9	99.2
	Skovgaard's $w^*(\psi)$, (2)	1.2	3.0	5.8	11.1	26.2	50.9	75.4	90.2	95.1	97.6	99.0
	Directional, (11)	1.2	2.9	5.8	10.9	25.8	50.3	75.0	89.9	95.0	97.5	99.0
	Standard error	0.0	0.0	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.0

4.3 Binary regression

Consider the data on page 249 of Andrews and Herzberg (1985) concerning calcium oxalate crystals in samples of urine. The binary response is an indicator of the presence of such crystals, and there are six explanatory variables: specific gravity, i.e., the density of urine relative to water; pH (ph); osmolarity (mOsm); conductivity (mMho); urea concentration (millimoles per litre); and calcium concentration (millimoles per litre). In the following analysis we use the $n = 77$ complete observations. A natural starting point for analysis is a logistic regression model with

$$\Pr(y_i = 1) = \mu_i(\theta) = \frac{\exp(x_i^\top \theta)}{1 + \exp(x_i^\top \theta)}, \quad i = 1, \ldots, n,$$

where x_i represents the vector of explanatory variables associated with the ith response y_i. The log-likelihood is of linear exponential form with canonical parameter $\varphi = \theta$, i.e.,

$$\ell(\varphi; y) = \varphi^\top X^\top y - 1_n^\top \log\{1_n + \exp(X\varphi)\},$$

where $y = (y_1, \ldots, y_n)^\top$ and X is the matrix of explanatory variables, with ith row $x_i^\top$.

The development of the directional p-value is similar to that for contingency tables in §4.1. In particular, to compute $\hat{\varphi}(t)$ we again solve equation (15) through iterative weighted least squares, but now with $\mu(\varphi) = \exp(X\varphi)/\{1+\exp(X\varphi)\}$. In this case the largest admissible value $t_{\max}$ is the largest value of t for which all fitted probabilities $\hat{\mu}(t)$ are non-negative and not larger than 1. Function $h(t; \psi)$ in (11), given by (10), is then

$$\begin{aligned} h(t;\psi) &= \exp\left[\hat{\mu}(t)^\top \left\{\log \hat{\mu}_\psi^0 - \log \hat{\mu}(t)\right\} + \{1_n - \hat{\mu}(t)\}^\top \left\{\log(1_n - \hat{\mu}_\psi^0) - \log(1_n - \hat{\mu}(t))\right\}\right] \\ &\quad \times |X^\top \mathrm{diag}\{\hat{\mu}(t)(1 - \hat{\mu}(t))\} X|^{-1/2}. \end{aligned}$$

For illustration, we compare a smaller model with the three covariates pH, osmolarity and conductivity to a full model with all six covariates, as in the formulation of Brazzale et al. (2007, p. 42); there are four nuisance parameters and three interest parameters. The directional p-value (11) for testing equivalence of the two models is 0.010, while the p-value from the χ^2_3 approximation to the log-likelihood ratio test is

0.004, and to Skovgaard's w^* is 0.011. Brazzale and Davison (2008, §4.2) discuss why higher-order corrections may be expected to be large in binary response models.

Inference for vector parameters is often needed when one or more covariates are factor variables with several levels, as the natural hypothesis of interest is that the factor variable has no effect on the response. As an example, we use the bacteria data from Venables and Ripley (2002, §10.4), which has a binary response, presence/absence of bacteria, and measurements on 50 subjects at 5 times. There are just 24 subjects that are informative for ψ, and 108 observations, an average of 4.5 observation per subject. The parameter of interest is ψ is a 5-level factor variable for time, and the nuisance parameters are the 24 subject-specific intercepts. Venables and Ripley (2002) used this example to illustrate the use of conditional likelihood with large numbers of nuisance parameters. The test of the hypothesis that the four between-week contrasts are all zero using the likelihood ratio statistic gave a p-value of 0.0005. The more accurate directional test gave a much larger p-value of 0.0054, Skovgaard's w^* gave a p-value of 0.0043. The exact conditional p-value is 0.0038; the difference between this and the two higher-order approximations is due to approximating a discrete distribution by a continuous one.

5 Examples with continuous response

5.1 Comparison of normal variances

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 variances among the g groups

$$H_0 : \sigma_1^2 = \cdots = \sigma_g^2$$

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

The model is a full exponential family and the log-likelihood for the parameter $\theta = (\mu_1, \ldots, \mu_g, \sigma_1^2, \ldots, \sigma_g^2)$ is

$$\ell(\theta; y) = -\frac{1}{2} \sum_{i=1}^{g} \left\{ n_i \log \sigma_i^2 + \frac{1}{\sigma_i^2} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 \right\} . \tag{16}$$

The full and the constrained maximum likelihood estimates are respectively

$$\hat{\theta} = (\bar{y}_1, \ldots, \bar{y}_g, v_1^2, \ldots, v_g^2)\,, \quad \hat{\theta}_0 = (\bar{y}_1, \ldots, \bar{y}_g, \bar{v}^2, \ldots, \bar{v}^2)\,,$$

where $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$, $v_i^2 = n_i^{-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ and $\bar{v}^2 = \sum_{i=1}^{g} n_i v_i^2 / \sum_{i=1}^{g} n_i$. Hence the log-likelihood ratio statistic is

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

which follows asymptotically the χ^2_{g-1} distribution, under the null hypothesis. The usual statistic for testing H_0 is due to Bartlett (1937),

$$\tilde{w} = \frac{\sum_{i=1}^{g} (n_i - 1) \log(\bar{s}^2 / s_i^2)}{1 + \{\sum_{i=1}^{g} (n_i - 1)^{-1} - (N - g)^{-1}\} \big/ \{3(g-1)\}^{-1}}\,,$$

where $N = \sum_{i=1}^{g} n_i$, $s_i^2 = n_i v_i^2 / (n_i - 1)$ and $\bar{s}^2 = N\bar{v}^2/(N - g)$. This is derived by Bartlett correction of the likelihood ratio statistic derived from the marginal likelihood for $\sigma_1^2, \ldots, \sigma_g^2$, based on the distribution of $s_1^2, \ldots, s_g^2$ (Barndorff-Nielsen and Cox, 1994, Example 6.16); see §6.

The model (16) 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 variances, $H_0 : \varphi_{j+1} - \varphi_j = 0, j = g+2, \ldots 2g$, is linear in the canonical parameter φ, and $K(\varphi) = -\sum_{i=1}^{g} n_i \{2\log(-2\varphi_{g+i}) + \varphi_i^2 \varphi_{g+i}^{-1}\}/4$. The global and constrained maximum likelihood estimates are

$$\hat{\varphi}^\top = \left(\frac{\bar{y}_1}{v_1^2}, \ldots, \frac{\bar{y}_g}{v_g^2}, -\frac{1}{2v_1^2}, \ldots, -\frac{1}{2v_g^2} \right), \tag{17}$$

$$\hat{\varphi}_0^\top = \left(\frac{\bar{y}_1}{\bar{v}^2}, \ldots, \frac{\bar{y}_g}{\bar{v}^2}, -\frac{1}{2\bar{v}^2}, \ldots, -\frac{1}{2\bar{v}^2} \right), \tag{18}$$

where for simplicity we write $\hat{\varphi}$ for $\hat{\varphi}^0$ and $\hat{\varphi}_0$ for $\hat{\varphi}^0_\psi$.

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

$$s_\psi = -\ell^0_\varphi(\hat{\varphi}_0) = \{0, \ldots, 0, -n_1(v_1^2 - \bar{v}^2), \ldots, -n_g(v_g^2 - \bar{v}^2)\}\,.$$

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 computed explicitly, giving

$$\begin{aligned}\ell(\varphi; t) &= \ell\{\varphi; s(t)\} \\ &= \sum_{i=1}^{g} n_i \left[\varphi_i \bar{y}_i + \varphi_{g+i}\left\{\bar{y}_i^2 + (tv_i^2 + (1-t)\bar{v}^2)\right\} + \frac{1}{2}\log(-2\varphi_{g+i}) + \frac{1}{4}\varphi_i^2\varphi_{g+i}^{-1}\right],\end{aligned}$$

which is maximized at

$$\hat{\varphi}_i(t) = \frac{\bar{y}_i}{tv_i^2 + (1-t)\bar{v}^2}, \quad \hat{\varphi}_{g+i}(t) = -\frac{1}{2\{tv_i^2 + (1-t)\bar{v}^2\}}\,, \quad i = 1, \ldots, g\,. \tag{19}$$

As expected, $t = 0$ and $t = 1$ give (18) and (17), respectively. Moreover, since $\hat{\varphi}_{g+i}(t)$ must be negative for all $i = 1, \ldots, g$, we have that

$$t < t_{\max} = \frac{\bar{v}^2}{\bar{v}^2 - \min_i v_i^2};$$

$s(t_{\max})$ is the last value of s along the line $s(t)$ that leads to an admissible maximum likelihood estimate (19). The directional p-value is computed from (11), with (10) giving

$$h(t) \propto \prod_{i=1}^{g} \{tv_i^2 + (1-t)\bar{v}^2\}^{(n_i-3)/2}\,. \tag{20}$$

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

$$\gamma = \left\{\sum_{i=1}^{g} \frac{n_i(v_i^2 - \bar{v}^2)^2}{\bar{v}^2}\right\}^{d/2} \left(\prod_{i=1}^{g} \frac{\bar{v}^2}{v_i^2}\right)^{3/2} \Big/ \left\{\frac{w}{2}\right\}^{d/2-1} \left\{\sum_{i=1}^{g} \frac{n_i(v_i^2 - \bar{v}^2)^2}{v_i^2\bar{v}^2}\right\}; \tag{21}$$

see the Supplementary Notes for more details.

Table 4: Data used to illustrate comparison of variances (NIST, 2012). Sufficient statistics for the gear diameter measurement of $g = 10$ batches each of $n = 10$ observations.

Batch	1	2	3	4	5	6	7	8	9	10
$10^2\bar{y}_i$	99.80	99.91	99.54	99.82	99.19	99.88	100.15	100.04	99.83	99.48
$10^5\hat{\sigma}_i^2$	1.70	2.45	1.42	1.34	5.17	8.80	5.59	1.18	1.54	2.56

When $g = 2$, so that the parameter of interest is scalar, and with equal group sizes $n_1 = n_2$, the directional p-value is identical to the p-value from the usual F-test. Such equality does not hold for $n_1 \neq n_2$, although simulations not given here indicate that the differences are slight. When $d = 1$, Skovgaard (2001)'s $w^* = r^{*2}$, which is very close numerically to the F-statistic, but not identical to it.

We illustrate these calculations using data on measurements of gear diameter for $g = 10$ batches of gears, with $n_i = 10$ observations from each batch. Summary statistics for the data are given in Table 4. The first order p-value based on the likelihood ratio statistic w is 0.0042; Bartlett's test gives a much larger p-value of 0.0136. The directional p-value 0.0389 is still larger, and Skovgaard's w^* gives a p-value of 0.0622. The pattern illustrated by these results is typical of the examples we have looked at; the first-order p-value seems to be too small, while w^* seems to over-correct.

We compared the accuracy of the approximations by simulation of balanced samples with varying numbers of groups, g, and observations per group, n. These were summarized by graphs that compare the p-values obtained from simulations under the hypothesis to the uniform distribution. For each configuration we considered 100,000 replications, with $\sigma_i^2 = 1$ and $\mu_i = 2(g - i)$ for $i = 1, \ldots, g$. The results are shown for two cases in Figure 3, with further results given in the Supplementary Notes. In the left panel, $g = 3$ and $n_i = 5$, giving two interest parameters and four nuisance parameters. In the right panel we took the extreme case of $g = 1000$ with $n_i = 5$; this has 999 interest parameters and 1001 nuisance parameters. As might be guessed from

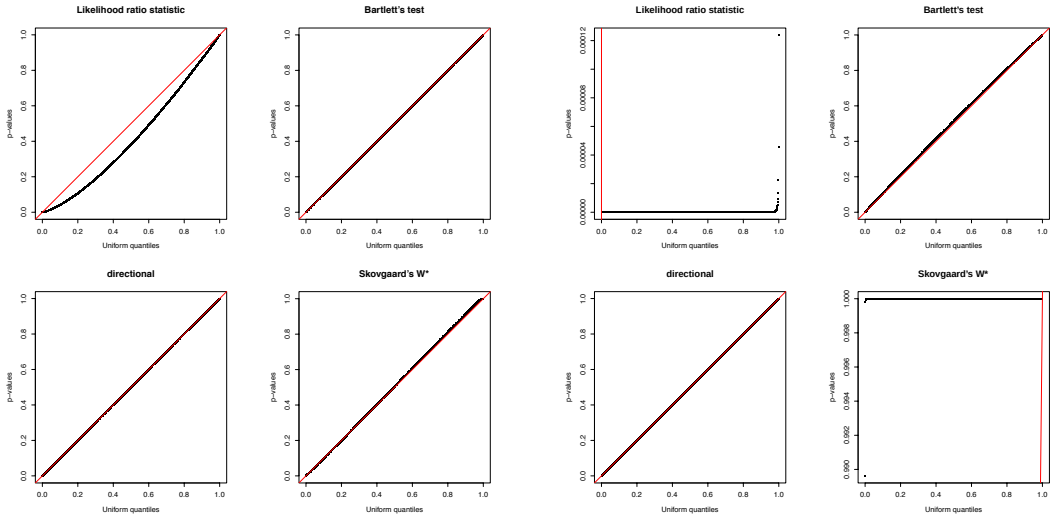


Figure 3: Simulations for testing common variances in $g = 3$ groups with $n_i = 5$ observations per group (left panels), and in $g = 1000$ groups with $n_i = 5$ observations per group (right panels), based on 100,000 replications. We compare the simulated p-values under the null hypothesis to the uniform distribution.

the gear data example, the likelihood ratio statistic yields p-values that are too small, but this is corrected by Bartlett's statistic $\tilde{w}$. The directional p-value is remarkably accurate in all cases, with a distribution practically indistinguishable from that from $\tilde{w}$, although the p-values in individual cases can be different. When $g = 1000$, inference based on the likelihood ratio statistic or on Skovgaard (2001)'s statistic w^* break down completely, but Bartlett's test and the directional test maintain their level extremely well; with a slight edge to the directional test apparent in Table S.2 in the Supplementary Notes. The more realistic cases of $g = 10$ and $n = 20$ are also reported in the Supplementary Notes: the likelihood ratio test and Skovgaard's statistic are noticeably non-uniform, whereas the directional test and Bartlett's test are essentially exact. We also computed an alternative version of w^*, $w^{**} = w - 2\log\gamma$, which is asymptotically equivalent to w^* in (2), but in all cases w^* outperforms w^{**}.

5.2 Comparison of exponential rates

In the model of the previous subsection, the dimensions of both interest and nuisance parameters increase with the number of groups, g. We now consider a model where the nuisance parameter is always scalar, although the dimension of the interest parameter increases. Suppose y_{ij} are independent random variables following an exponential distribution with rates θ_i, for $i = 1, \ldots, g$ and $j = 1, \ldots, n_i$. The hypothesis of interest is homogeneity of the rates among the g groups, $\theta_1 = \cdots = \theta_g$, the alternative being that at least one equality does not hold. The log-likelihood for the parameter $\theta = (\theta_1, \ldots, \theta_g)$ is

$$\ell(\theta; y) = \sum_{i=1}^{g} \{-u_i\theta_i + n_i \log \theta_i\} \ ,$$

where $u_i = n_i\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}$; the canonical parameter $\varphi = -\theta$ and the sufficient statistic is $u = (u_1, \ldots, u_g)$. The hypothesis can be expressed as a linear constraint on the canonical parameter, i.e.,

$$H_0 : \psi_1 = \cdots = \psi_{g-1} = 0 \, ,$$

with, for instance, $\psi_i = \theta_{i+1} - \theta_i$, for $i = 1, \ldots, g-1$.

The full and the constrained maximum likelihood estimates are respectively

$$\hat{\theta} = (\bar{y}_1^{-1}, \ldots, \bar{y}_g^{-1}) \, , \quad \hat{\theta}^0 = (\bar{y}^{-1}, \ldots, \bar{y}^{-1}) \, ,$$

where $\bar{y} = \sum_{i=1}^{g} n_i\bar{y}_i / \sum_{i=1}^{g} n_i$ and the log-likelihood ratio statistic is

$$w = 2\sum_{j=1}^{n_i} n_i \log(\hat{\theta}_i/\hat{\theta}^0) = 2\sum_{j=1}^{n_i} n_i \log(\bar{y}/\bar{y}_i) \tag{22}$$

which has an asymptotic χ^2_{g-1} distribution under the null hypothesis.

The tilted log-likelihood (6) along the line $s(t)$,

$$\begin{aligned} \ell(\varphi; t) &= \ell\{\varphi; s(t)\} = \ell^0(\varphi) + \varphi^\top s(t) \\ &= \sum_{i=1}^{g} [\{u_i + n_i(1-t)(\bar{y} - \bar{y}_i)\}\varphi_i + n_i \log(-\varphi_i)] \ , \end{aligned}$$

is maximized at

$$\hat{\varphi}_i(t) = -\frac{1}{\bar{y} - t(\bar{y} - \bar{y}_i)} \, , \quad i = 1, \ldots, g \, . \tag{23}$$

The line for the directional test goes through s^0 and $s_\psi = \{-n_1(\bar{y}_1 - \bar{y}), \ldots, -n_g(\bar{y}_g - \bar{y})\}$, where $\hat{\varphi}(0) = -\hat{\theta}^0$, and since $\hat{\varphi}_i(t)$ has to be negative for all $i = 1, \ldots, g$, we have that

$$t < t_{\max} = \frac{\bar{y}}{\bar{y} - \min_i \bar{y}_i} .$$

The directional p-value (11) uses $h(t)$ from (10)

$$h(t) \propto \prod_{i=1}^{g} \{1 - t(\bar{y} - \bar{y}_i)/\bar{y}\}^{(n_i - 1)} ,$$

since $|j_{\varphi\varphi}(\varphi; s)| = \prod_{i=1}^{g} n_i \varphi_i^{-2}$.

Skovgaard (2001)'s modification can again be computed explicitly, and is

$$\gamma = \left\{ \sum_{i=1}^{g} n_i \left(\frac{\bar{y}_i - \bar{y}}{\bar{y}} \right)^2 \right\}^{(g-1)/2} \left(\prod_{i=1}^{g} \frac{\bar{y}}{\bar{y}_i} \right) \Big/ \left\{ w^{(g-1)/2} \sum_{i=1}^{g} \frac{n_i (\bar{y}_i - \bar{y})^2}{\bar{y}\bar{y}_i} \right\} . \qquad (24)$$

We illustrate these calculations by testing the equality of the mean times between failures of the air-conditioning equipment in ten Boeing 720 aircraft (Proschan, 1963; Cox and Snell, 1981) The first order p-value based on (22) equals 0.0198, the directional p-value (11) equals 0.0227, and Skovgaard's modified likelihood ratio statistic (2) equals 0.0274.

Table 5 summarizes simulation studies using the same sample sizes as in the example, and in a balanced but more extreme setting. The results confirm the very accurate behaviour of the directional approach, while showing a worsening of the performance of both first order and Skovgaard's statistics in the second setting where the samples sizes are relatively small.

5.3 Covariance selection

A linear exponential model of interest in the analysis of graphical models concerns inference about entries of the concentration, or inverse covariance, matrix in a multivariate normal distribution. A zero entry in the concentration matrix implies conditional independence of two variables given the values of other variables and corresponds to no arc between nodes representing the two variables in a conditional independence graph (Lauritzen, 1996).

Table 5: Simulated empirical distribution (%) of p-values for testing equality of exponential rates, based on $100,000$ replications. The upper figures are for data with $g = 10$ groups with sample sizes $23, 29, 15, 14, 30, 27, 24, 9, 12, 16$, and the lower ones are for $g = 10$ groups and sample sizes $n_i = 5$.

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, (22)	1.1	2.7	5.4	10.6	25.8	50.7	75.5	90.3	95.2	97.6	99.0
Skovgaard's w^*, (24)	0.9	2.4	4.7	9.6	23.9	48.2	73.1	88.7	94.2	96.9	98.6
Directional, (11)	1.0	2.6	5.0	10.2	25.0	49.9	74.9	90.0	95.0	97.5	99.0
Likelihood ratio, (22)	1.3	3.1	5.9	11.5	27.5	52.9	77.0	91.0	95.5	97.7	99.0
Skovgaard's w^*, (24)	0.8	2.0	3.9	8.0	21.0	44.1	68.8	85.5	91.8	95.3	97.7
Directional, (11)	1.0	2.5	4.9	9.9	24.8	50.0	75.0	89.9	94.9	97.5	98.9
Standard error	0.0	0.0	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.0

Let $y_1, \ldots, y_n$ be a sample of independent random vectors from a multivariate normal $N_q(\mu, \Lambda^{-1})$, where the mean μ and the concentration matrix Λ are unknown and arbitrary apart from the restriction that Λ is positive definite. Let y denote the $n \times q$ matrix with ith row vector $y_i^\top$. Then the log-likelihood for $\theta = (\mu, \Lambda)$ is

$$\ell(\theta; y) = \frac{n}{2} \log |\Lambda| - \frac{1}{2} \mathrm{tr}(\Lambda y^\top y) + 1_n^\top y \, \Lambda \mu - \frac{n}{2} \mu^\top \Lambda \mu \, .$$

This model is saturated and the maximum likelihood estimate exists if and only if the matrix $y^\top y - y^\top 1_n 1_n^\top y/n$ is positive definite, which happens with probability one if $n > q$ (Lauritzen, 1996, Theorem 5.1). The maximum likelihood estimate $\hat{\theta}$ has components

$$\hat{\mu} = y^\top 1_n / n, \qquad \hat{\Lambda}^{-1} = y^\top y / n - y^\top 1_n 1_n^\top y / n^2 \, .$$

Consider now a reduced model in which some off-diagonal elements of Λ equal zero. With ψ denoting the $d \times 1$ vector of these components, the reduced model corresponds to the null hypothesis $H_0 : \psi = 0$. Under H_0 the constrained maximum likelihood estimate of θ is $\hat{\theta}_0 = (\hat{\mu}, \hat{\Lambda}_0)$, where $\hat{\Lambda}_0$ is typically obtained numerically, for instance using the `R` function `fitConGraph` in package `ggm`, and as $n \to \infty$ the log-likelihood ratio statistic,

$$w = -n \log(|\hat{\Lambda}^{-1} \hat{\Lambda}_0|) \, , \tag{25}$$

follows the χ^2_d distribution.

The canonical parameter for this exponential family is $\varphi = (\xi, \Lambda) = (\Lambda\mu, \Lambda)$, with corresponding log-likelihood

$$\ell(\varphi; y) = \frac{n}{2} \log |\Lambda| - \frac{1}{2} \mathrm{tr}(\Lambda y^\top y) + 1_n^\top y \, \xi - \frac{n}{2} \xi^\top \Lambda^{-1} \xi \, . \tag{26}$$

The expected value s_ψ defined in (8) is $s_\psi = -\{\ell_\xi(\hat{\varphi}_0), \ell_\Lambda(\hat{\varphi}_0)\} = \{0, n(\hat{\Lambda}^{-1} - \hat{\Lambda}_0^{-1})/2\}$.The tilted log-likelihood (6) along the line $s(t) = (1-t)s_\psi$ can be obtained using (26). The maximization is straightforward in the θ parameterization and yields $\hat{\theta}(t) = \{\hat{\mu}, \hat{\Lambda}(t)\}$, with $\hat{\Lambda}(t)^{-1} = t\hat{\Lambda}^{-1} + (1-t)\hat{\Lambda}_0^{-1}$. The last value of s along the line $s(t)$, $s(t_{\max})$, is the largest value such that $\hat{\Lambda}(t)$ is positive definite, and this can easily be found numerically.

The directional p-value (11) uses $h(t)$ from (10), and since $\left|\jmath_{\varphi\varphi}\left[\hat{\varphi}\{s(t)\}; s(t)\right]\right|^{-1/2} = |\hat{\Lambda}(t)|^{(q+2)/2}$, we find that

$$h(t) \propto |\hat{\Lambda}(t)|^{-(n-q-2)/2} \propto |t\hat{\Lambda}^{-1} + (1-t)\hat{\Lambda}_0^{-1}|^{(n-q-2)/2}.$$

In this example Skovgaard (2001)'s modified likelihood ratio statistic (2) has

$$\gamma = \left\{\frac{1}{2}\left[\mathrm{tr}(\hat{\Lambda}^{-1}\hat{\Lambda}_0\hat{\Lambda}^{-1}\hat{\Lambda}_0) - q\right]\right\}^{d/2} |\hat{\Lambda}^{-1}\hat{\Lambda}_0|^{-(q+2)/2} \Big/ \frac{1}{2}\left\{\mathrm{tr}(\hat{\Lambda}\hat{\Lambda}_0^{-1}) - q\right\}\left\{-\log|\hat{\Lambda}^{-1}\hat{\Lambda}_0|\right\}^{d/2-1}. \tag{27}$$

We illustrate this model using the dataset of Kenward (1987, Table 1), which consists of repeated measurements of weights (kg) of 60 calves from a trial on the control of intestinal parasites. The animals were put out to pasture at the start of the grazing season, and each was then weighed on 11 occasions. The first ten measurements were made at two-weekly intervals, with a final one made after a further week. We test first-order Markovian dependence of the measurements, i.e., we test that all off-diagonal elements of Λ are zero, except those closest to the diagonal. In the saturated model Λ has 66 parameters, while in the reduced model it has 21 parameters, so $d = 45$. The log-likelihood ratio statistic is $w = 68.377$ and gives p-value 0.0139 based on its asymptotic χ^2_{45} distribution. The directional p-value is 0.0706, while Skovgaard's $w^* = 57.243$, with p-value 0.1042.

The upper part of Table 6 summarizes a simulation study from the fitted reduced model. The results underline the high accuracy of the directional approach, while the performances of the first order and Skovgaard's statistics are respectively poor and not very accurate. To explore how robust this was to the dimension, we considered much larger matrices, with $q = 30$ and 50, giving likelihood ratio tests with 406 and 1176 degrees of freedom respectively, the last two approaches are catastrophically bad, but the directional approach retains its excellent performance.

Inference on covariance matrices in the multivariate normal are sometimes based on the Wishart marginal distribution of the sample covariance matrix, which is free of of the nuisance parameters μ; in some contexts this is called the restricted likelihood function, or REML. The directional p-value obtained using this marginal distribution is identical to the one developed above starting from the full likelihood (26). The

Table 6: Simulated empirical distribution (%) of p-values for testing first-order Markov dependence with $n = 60$, based on $100,000$ replications. The dimension q of the covariance matrix is 11, 30 and 50 for the top, middle and lower rows, respectively; the dimension of the parameter of interest is correspondingly 45, 406 and 1176.

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, (25)	5.5	10.5	17.0	27.0	48.7	73.0	89.5	96.7	98.5	99.4	99.8
Skovgaard's w^*, (27)	0.7	1.8	3.6	7.4	19.6	42.2	67.8	85.2	91.9	95.5	98.0
Directional, (11)	1.1	2.6	5.0	10.1	24.8	49.8	74.9	89.9	94.9	97.4	99.0
Likelihood ratio, (25)	91.2	95.4	97.5	98.9	99.8	100	100	100	100	100	100
Skovgaard's w^*, (27)	0.0	0.0	0.0	0.2	1.1	4.7	14.8	31.5	44.0	55.6	68.5
Directional, (11)	1.0	2.5	5.0	10.1	25.2	50.2	75.1	90.1	95.0	97.5	99.0
Likelihood ratio, (25)	100	100	100	100	100	100	100	100	100	100	100
Skovgaard's w^*, (27)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.7	1.6	2.9	5.6
Directional, (11)	1.0	2.5	5.0	10.0	25.0	49.8	74.8	89.9	94.9	97.5	99.0
Standard error	0.0	0.0	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.0

p-values using w and w^* would be slightly different, although simulation results not shown here indicate that numerically there is no practical difference in using the full or the marginal likelihoods to compute w and w^*.

6 Discussion

We have presented the formulas for the conditional density $h(t)$ on the line, (10), and associated p-value, (11), in the context of inference for linear functions of the canonical parameter in an exponential family. Elimination of nuisance parameters by conditioning is only available in this setting. To construct a directional test for inference on nonlinear functions of the canonical parameter we first need a reference density analogous to (7). Fraser (2012) shows that this reference distribution is a marginal density for a particular derived variable, and that the saddlepoint approximation to this density has a form similar to (7), but with an additional adjustment for curvature.

More generally, if the underlying model is not an exponential family, then the method may be extended by first approximating the model by a so-called *tangent* exponential family. This entails constructing a nominal canonical parameter $\varphi(\theta)$ from the original model, using arguments built on approximate ancillarity. The tangent exponential family was used to develop r^*-type approximations in Fraser and Reid (1995); see also Fraser et al. (1999). An overview is given in Brazzale et al. (2007, Ch. 8).

The chi-squared approximation to the distribution of $w(\psi)$ can be improved by Bartlett correction (Bartlett, 1937); it can be shown that

$$\tilde{w}(\psi) = w(\psi)/[\mathrm{E}_\theta\{w(\psi)\}/d] \tag{28}$$

follows a χ^2_d distribution with relative error $O(n^{-2})$. Skovgaard (2001) notes that the accuracy of the χ^2_d approximation to (28) can be lost when the expected value is approximated using its asymptotic expansion, rather than computed analytically. Even the approximate version can be cumbersome to compute, as it involves arrays of third and fourth order cumulants (Lawley, 1956; McCullagh and Cox, 1986). The comparison of normal variances in §5.1 is exceptional in that an analytical expression

for the Bartlett correction is available, although in that case the likelihood that is corrected is the marginal likelihood for the variances, which already has an adjustment to the degrees of freedom. The directional test implements this degrees of freedom adjustment automatically, via the saddlepoint approximation. The Bartlett test, like the likelihood ratio test, is an 'omnibus' test, looking in all directions of the parameter space for alternatives. In the scalar parameter setting, this means that error may be larger than the nominal in one tail of the distribution, and smaller than nominal in the other. The directional test, on the other hand, looks in the direction determined by the data.

Directional tests have been proposed before, but have not been widely used. Skovgaard (1988) and Cheah et al. (1994) attempted to avoid numerical integration by using an integration by parts argument analogous to that yielding r^*. However, the presence of the Jacobian term t^{d-1} in the density means that the base distribution for this integration by parts is χ^2_d, rather than standard normal, and for reasons that are unclear the approximation is not nearly as accurate as the normal approximation to the distribution of r^*, a phenomenon also noted by Wood et al. (1993). Also previously overlooked was the simplification of the conditional density when evaluated only on the line $\mathcal{L}^*$, where any factors not involving t cancel from the numerator and the denominator.

In all the examples treated here, the directional p-value can be computed in `R` (R Development Core Team, 2012) by first fitting a full and a constrained generalized linear model using `glm`, and then computing the one-dimensional integral with `integrate`. The only non-standard aspect is the determination of $t_{\max}$ in (11). For contingency tables, as discussed in §4.1, $t_{\max}$ can be obtained explicitly if the hypothesis is nested in the saturated model. If the hypothesis is nested in an unsaturated model, as in the last example in §4.1, then $t_{\max}$ is reached when margins of certain subtables are zero; a general treatment is given in Fienberg and Rinaldo (2012). Our implementation for cases where $t_{\max}$ is not available explicitly simply fits the model for increasing values of t until the maximum likelihood estimate reaches the boundary of the parameter space.

For some of the contingency table examples in §4, algorithms are available to compute the exact p-value, conditional on the table margins. The commercial package

StatXact (Mehta, 1991) uses a network algorithm for this computation, but for larger sample sizes some type of sampling is usually needed. The `R` package `exactLoglinTest` (Caffo, 2006) uses either importance sampling or Markov chain sampling; both are built on a normal approximation to the Poisson distribution. This package can be used to test independence, although we found in applying it to the data of Table 2 that careful tuning of the algorithm was needed. Conditional simulation can also be implemented with the Metropolis–Hastings algorithm (Diaconis and Sturmfels, 1998; Forster et al., 1996, 2003; Smith et al., 1996), but ensuring irreducibility of the resulting chain is not straightforward in general, and so far as we know no general code is available for this. Caffo and Booth (2003) gives a helpful overview of Monte Carlo methods for log linear models.

Unconditional simulation from the fitted model would be expected to give lower theoretical accuracy than the approach described above, and although precision can be improved by nested simulation (Davison and Hinkley, 1997, §4.5) the computational burden would then greatly increase. An unconditional approach is proposed in DiCiccio and Young (2008), but it seems to be available only for scalar parameters of interest.

The balance between mathematical elegance and computational brute force is a matter of taste, but even practical considerations suggest that the demonstrated accuracy of the directional approach makes it worthy of broad consideration. It has the added advantage that the same method can be used in continuous models.

References

Agresti, A. (2002). *Categorical Data Analysis* (Second ed.). New York: John Wiley.

Andrews, D. F. and A. M. Herzberg (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. New York: Springer.

Barndorff-Nielsen, O. E. and D. R. Cox (1979). Edgeworth and saddle-point approximations with statistical applications (with Discussion). *J. R. Statist. Soc. B 41*, 279–312.

Barndorff-Nielsen, O. E. and D. R. Cox (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.

Brazzale, A. R. and A. C. Davison (2008). Accurate parametric inference for small samples. *Statistical Science 23*, 465–484.

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

Butler, R. W. (2007). *Saddlepoint Approximations with Applications*. Cambridge: Cambridge University Press.

Caffo, B. (2006). Exact hypothesis test for log-linear models with `exactLoglinTest`. *Journal of Statistical Software 17*(7).

Caffo, B. and J. Booth (2003). Monte carlo conditional inference for a log-linear and logistic models: a survey of current methodology. *Statistical Methods in Medical Research 12*, 109–123.

Cheah, P., D. A. S. Fraser, and N. Reid (1994). Multiparameter testing in exponential models: third order approximations from likelihood. *Biometrika 81*, 271–278.

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

Cox, D. R. and E. J. Snell (1981). *Applied Statistics: Principles and Examples*. London: Chapman & Hall.

Davison, A. C. (1988). Approximate conditional inference in generalized linear models. *J. R. Statist. Soc. B 50*, 445–461.

Davison, A. C., D. A. S. Fraser, and N. Reid (2006). Improved likelihood inference for discrete data. *J. R. Statist. Soc. B 68*, 495–508.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Diaconis, P. and B. Sturmfels (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics 26*, 363–397.

DiCiccio, T. J. and G. A. Young (2008). Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families. *Biometrika 95*, 747–758.

Everitt, B. S. (1992). *The Analysis of Contingency Tables* (Second ed.). London: Chapman & Hall.

Fienberg, S. E. and A. Rinaldo (2012). Maximum likelihood estimation in log-linear models. *Ann. Statist. 40*, 996–1023.

Forster, J. J., J. W. McDonald, and P. W. F. Smith (1996). Monte Carlo exact conditional tests for log-linear and logistic models. *Journal of the Royal Statistical Society, series B 58*, 445–453.

Forster, J. J., J. W. McDonald, and P. W. F. Smith (2003). Markov chain Monte Carlo exact inference for binomial and multinomial regression models. *Statistics and Computing 13*, 169–177.

Fraser, D. A. S. (2012). Assessing an interest parameter: the data and a definitive reference distribution. submitted for publication.

Fraser, D. A. S. and H. Massam (1985). Conical tests: Observed levels of significance and confidence regions. *Statist. Hefte 26*, 1–17.

Fraser, D. A. S. and N. Reid (1995). Ancillaries and third order significance. *Utilitas Mathematica 47*, 33–53.

Fraser, D. A. S. and N. Reid (2006). Assessing a vector parameter. *Student 5*, 247–256.

Fraser, D. A. S., N. Reid, and J. Wu (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika 86*, 249–264.

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *J. R. Statist. Soc. C 36*, 296–308.

Kolassa, J. E. and M. A. Tanner (1994). Approximate conditional inference in exponential families via the Gibbs sampler. *J. Am. Statist. Assoc. 89*, 697–702.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Oxford University Press.

Lawley, D. N. (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika 43*, 295–303.

McCullagh, P. and D. R. Cox (1986). Invariants and likelihood ratio statistics. *Ann. Statist. 14*, 1419–1430.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (Second ed.). London: Chapman & Hall.

Mehta, C. (1991). Statxact: A statistical package for exact nonparametric inference. *The American Statistician 45*, 74–75.

NIST (2012, April). NIST/SEMATECH e-Handbook of Statistical Methods. `http://www.itl.nist.gov/div898/handbook/eda/section3/eda357.htm`. Accessed October 19, 2012.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics 5*, 375–383.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

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.

Smith, P. W. F., J. J. Forster, and J. W. McDonald (1996). Monte Carlo exact tests for square contingency tables. *Journal of the Royal Statistical Society, series A 159*, 309–321.

Venables, W. and B. Ripley (2002). *Modern Applied Statistics with* `S` (4th ed.). New York: Springer.

Wood, A. T. A., J. G. Booth, and R. W. Butler (1993). Saddlepoint approximation to the CDF of some statistics with non-normal limit distributions. *J. Am. Statist. Assoc. 88*, 680–686.