Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Accurate Directional Inference for Vector Parameters in Linear Exponential Families

A. C. Davison, D. A. S. Fraser, N. Reid & N. Sartori

Published online: 19 Mar 2014.

To cite this article: A. C. Davison, D. A. S. Fraser, N. Reid & N. Sartori (2014) Accurate Directional Inference for Vector Parameters in Linear Exponential Families, Journal of the American Statistical Association, 109:505, 302-314, DOI: 10.1080/01621459.2013.839451

To link to this article: http://dx.doi.org/10.1080/01621459.2013.839451

Accurate Directional Inference for Vector Parameters in Linear Exponential Families

A. C. DAVISON, D. A. S. FRASER, N. REID, and N. SARTORI

We consider inference on a vector-valued parameter of interest in a linear exponential family, 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. The work simplifies and develops earlier research on directional tests for continuous models and on higher-order inference for discrete models, and the examples include contingency tables and logistic regression. Examples and simulations illustrate the high accuracy of the method, which we compare with the usual likelihood ratio test and with an adjusted version due to Skovgaard. In high-dimensional settings, such as covariance selection, the approach works essentially perfectly, whereas its competitors can fail catastrophically.

KEY WORDS: Contingency table; Covariance selection; Exponential family model; Higher-order asymptotics; Likelihood ratio test; Logistic regression

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. It also motivates the construction of test statistics in many other settings, such as empirical and semiparametric likelihood inference. In large samples, chi-squared approximations to the distribution of the likelihood ratio statistic may be used, but their accuracy may be poor, for example, in sparse contingency tables or high-dimensional graphical models. Thus, it is of wide interest to consider potentially more accurate approximations.

In this article, we discuss a directional approach derived from higher-order approximations for likelihood inference on vector parameters of interest. 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 center 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, Davison, and Reid (2007) and Brazzale and Davison (2008); the discrete case is considered in more generality in Davison, Fraser, and Reid (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 article provides an alternative approach for linear exponential family models, which improves on the usual likelihood ratio statistic, seems to be more accurate in simulations than Skovgaard's statistic, and is very easy to compute.

Directional tests for vector parameters of interest were proposed by Fraser and Massam (1985) and Skovgaard (1988). 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. For exponential family models, Cheah, Fraser, and Reid (1994) proposed using the sufficient statistic for the parameter as the vector measure of departure, and adapted the one-dimensional approximation to the vector setting. That article and Skovgaard (1988) used integration by parts arguments analogous to that yielding r^*, though with a χ^2_d, rather than standard normal, base distribution, but for reasons that are unclear the resulting approximation is much less accurate than the normal approximation to the distribution of r^*, a phenomenon also noted by Wood, Booth, and Butler (1993). The approach described here simplifies these earlier proposals and computes the p-value by one-dimensional numerical integration, evaluated conditionally on the direction of the variable; see (8) and (10). This is both computationally fast and very accurate, thus allowing the routine use of directional tests in practice.

In this article, we restrict attention to exponential family models in which 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 how they are likely to perform when extended to more general settings. As examples we consider multidimensional 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 our proposed approach is shown to be extremely accurate, even

Anthony Davison is Professor of Statistics, EPFL-FSB-MATHAA-STAT, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland (E-mail: *Anthony.Davison@epfl.ch*). Nancy Reid (E-mail: *reid@utstat.toronto.edu*) and Don Fraser (E-mail: *dfraser@utstat.toronto.edu*) are Professors of Statistics, Department of Statistics, University of Toronto, Toronto, Canada M5S 3G3. Nicola Sartori is Associate Professor of Statistics, Dipartimento di Scienze Statistiche, Università degli Studi di Padova, 35121 Padova, Italy (E-mail: *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 2011/2012. We thank the reviewers for their cordial and constructive comments.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

Journal of the American Statistical Association
March 2014, Vol. 109, No. 505, Theory and Methods
DOI: 10.1080/01621459.2013.839451

when examined unconditionally, and also in high-dimensional settings, where its competitors break down. The method also captures the structure of the models, for example, reproducing the F-test for comparing two variances in Section 5.1.

2. BACKGROUND

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

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

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

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

With relative error $O(n^{-1})$, $w(\boldsymbol{\psi})$ follows a χ^2_d distribution, with degrees of freedom d equal to the number of constrained parameters in $H_{\boldsymbol{\psi}}$. 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) attributed 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 proposed an analogous version for vector interest parameters designed to maintain accuracy in the tails of the distribution. The resulting quantity

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

uses a correction factor γ that compares w to an asymptotically equivalent quadratic form. Skovgaard (2001) showed that in addition to having good large-deviation properties, $w^*(\boldsymbol{\psi})$ is also easier to calculate than the Bartlett adjustment discussed in Section 6. This approach may be applied beyond the linear exponential families considered in the present article.

Like the likelihood ratio $w(\boldsymbol{\psi})$, Equation (2) gives an omnibus measure of departure; all potential directions away from the hypothesis $H_{\boldsymbol{\psi}}$ are averaged in the calculation of the p-value. We propose a measure of departure that incorporates information in the data about the relevant direction of deviation from $H_{\boldsymbol{\psi}}$, 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 $\boldsymbol{\lambda}$ is four-dimensional,

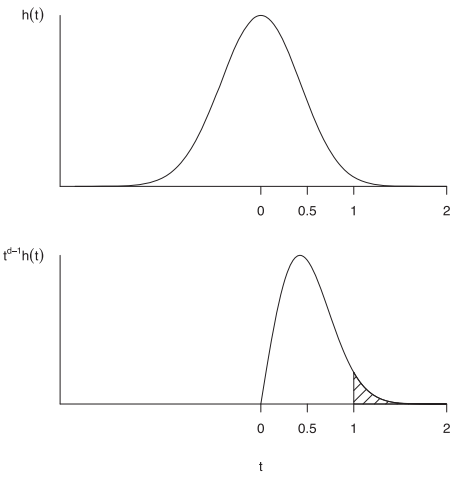
Table 1. Retarded activity among psychiatric patients (Everitt 1992, Table 3.3)

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

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

We measure departure from $H_{\boldsymbol{\psi}}$ 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 $h(t;\boldsymbol{\psi})$, given in (8), on this line, with points $t = 0, 0.5, 1, 2$ indicated. The lower left panel shows the shape of the relative density $t^{d-1}h(t;\boldsymbol{\psi})$, for $t > 0$, used in (10) to compute the directional p-value.

The directional p-value (10) 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's (2001) 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 p-values to differ by much.



$t = 0$

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 $h(t;\boldsymbol{\psi})$ on the line indexed by t (top left) and the directed radial distance from the value expected under independence, $t^{d-1}h(t;\boldsymbol{\psi})$ (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 ($t = t_{\max} = 2$).

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

3. DIRECTIONAL TESTS IN LINEAR EXPONENTIAL FAMILIES

3.1 Dimension Reduction by Conditioning

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

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

with associated log-likelihood function $\ell(\boldsymbol{\theta};\boldsymbol{y}) = \boldsymbol{\varphi}(\boldsymbol{\theta})^\top \boldsymbol{u}(\boldsymbol{y}) - K\{\boldsymbol{\varphi}(\boldsymbol{\theta})\}$, where in all log-likelihood functions we ignore additive terms that do not depend on $\boldsymbol{\theta}$. Since $\boldsymbol{u}$ is sufficient for $\boldsymbol{\varphi}$, the marginal density for $\boldsymbol{u}$ is of the same form, $f(\boldsymbol{u};\boldsymbol{\theta}) = \exp[\boldsymbol{\varphi}(\boldsymbol{\theta})^\top \boldsymbol{u} - K\{\boldsymbol{\varphi}(\boldsymbol{\theta})\}]\tilde{h}(\boldsymbol{u})$, and the log-likelihood function is $\ell(\boldsymbol{\theta};\boldsymbol{u}) = \boldsymbol{\varphi}(\boldsymbol{\theta})^\top \boldsymbol{u} - K\{\boldsymbol{\varphi}(\boldsymbol{\theta})\}$.

It is convenient in what follows to center the sufficient statistics at the observed data point $\boldsymbol{u}^0 = \boldsymbol{u}(\boldsymbol{y}^0)$, so we let $\boldsymbol{s} = \boldsymbol{u} - \boldsymbol{u}^0$, and write

$$\ell(\boldsymbol{\theta};\boldsymbol{s}) = \boldsymbol{\varphi}(\boldsymbol{\theta})^\top \boldsymbol{s} + \ell^0(\boldsymbol{\theta}) = \boldsymbol{\varphi}(\boldsymbol{\theta})^\top(\boldsymbol{u} - \boldsymbol{u}^0) + \ell(\boldsymbol{\theta};\boldsymbol{u}^0), \quad (3)$$

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

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

$$\ell(\boldsymbol{\varphi};\boldsymbol{s}) = \boldsymbol{\psi}^\top \boldsymbol{s}_1 + \boldsymbol{\lambda}^\top \boldsymbol{s}_2 + \ell^0(\boldsymbol{\psi}, \boldsymbol{\lambda}), \quad (4)$$

where $\boldsymbol{\psi}$ and $\boldsymbol{s}_1$ are d-dimensional. In this simpler model, the conditional distribution of $\boldsymbol{s}_1$, given $\boldsymbol{s}_2$, depends only on $\boldsymbol{\psi}$, and also has exponential family form:

$$f(\boldsymbol{s}_1 \mid \boldsymbol{s}_2;\boldsymbol{\psi}) = \exp\{\psi^\top \boldsymbol{s}_1 - \tilde{K}(\boldsymbol{\psi})\}\tilde{h}_2(\boldsymbol{s}_1), \quad (5)$$

although $\tilde{K}(\cdot)$ and $\tilde{h}_2(\cdot)$ can rarely be computed explicitly. Conditioning on the observed value of $\boldsymbol{s}_2$ is equivalent to fixing the constrained maximum likelihood estimate under $H_{\boldsymbol{\psi}}$, which is $\widehat{\boldsymbol{\varphi}}_{\boldsymbol{\psi}} = (\boldsymbol{\psi}, \widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}})$, and when computed at $\boldsymbol{y}^0$ is denoted by $\widehat{\boldsymbol{\varphi}}^0_{\boldsymbol{\psi}}$.

The saddlepoint approximation to this conditional density can be expressed as

$$\dot{f}(\boldsymbol{s};\boldsymbol{\psi})\,d\boldsymbol{s} = c\exp\left[\ell\left(\widehat{\boldsymbol{\varphi}}^0_{\boldsymbol{\psi}};\boldsymbol{s}\right) - \ell\{\widehat{\boldsymbol{\varphi}}(\boldsymbol{s});\boldsymbol{s}\}\right] \times |\boldsymbol{J}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}\{\widehat{\boldsymbol{\varphi}}(\boldsymbol{s});\boldsymbol{s}\}|^{-1/2}\,d\boldsymbol{s}, \quad \boldsymbol{s} \in \mathcal{L}^0, \quad (6)$$

where c normalizes the conditional density and $\mathcal{L}^0$ is a d-dimensional plane defined by setting $\boldsymbol{s}_2 = \boldsymbol{0}$, or equivalently setting $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}} = \widehat{\boldsymbol{\lambda}}^0_{\boldsymbol{\psi}}$. In (6), $\widehat{\boldsymbol{\varphi}}(\boldsymbol{s})$ is the solution in $\boldsymbol{\varphi}$ of the score equation from (4); $\boldsymbol{s} = -\ell^0_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) = -\partial\ell^0(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}$, and $\boldsymbol{J}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}(\boldsymbol{\varphi};\boldsymbol{s}) = -\partial^2\ell(\boldsymbol{\varphi};\boldsymbol{s})/\partial\boldsymbol{\varphi}\partial\boldsymbol{\varphi}^\top$.

Although it is more conventional to write $\dot{f}(\boldsymbol{s}_1 \mid \boldsymbol{s}_2;\boldsymbol{\psi})$ 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, that is, constraining $\boldsymbol{s}$ 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 is discussed in Davison (1988), though it was introduced, in the context of binomial data, by Daniels (1958). Butler (2007) gave a comprehensive review of saddlepoint approximations and their statistical applications. A direct derivation and presentation entirely in terms of likelihood, as above, is given in Fraser and Reid (2014), and generalized there to inference for nonlinear functions of the canonical parameters and to approximate exponential models.

3.2 Directional Departure

The density (6) on the d-dimensional subspace $\mathcal{L}^0$ is the basic density for inference about ψ, and leads to unbiased tests that are similar on the boundary in the usual way (Lehmann and Romano 2005, chap. 4): the nuisance parameter is eliminated from the model by conditioning. This conditional model can be used, for example, to derive a likelihood ratio test of H_ψ; the asymptotic distribution of the log-likelihood ratio statistic is χ^2_d, under the usual regularity conditions.

For more accurate p-values, we develop instead a one-dimensional model by considering the magnitude of $\boldsymbol{s}$, conditional on its direction. To this end we examine the line $\mathcal{L}^*$, in $\mathcal{L}^0$, obtained by joining the observed value of $\boldsymbol{s}$, which is $\boldsymbol{s}^0 = \boldsymbol{0}$, and the value $\boldsymbol{s}_{\boldsymbol{\psi}}$, which is the value of $\boldsymbol{s}$ for which $\boldsymbol{\varphi} = \widehat{\boldsymbol{\varphi}}^0_{\boldsymbol{\psi}}$ is the maximum likelihood estimate; from (4)

$$\boldsymbol{s}_{\boldsymbol{\psi}} = -\ell^0_{\boldsymbol{\varphi}}(\widehat{\boldsymbol{\varphi}}^0_{\boldsymbol{\psi}}) = \begin{bmatrix} -\ell^0_{\boldsymbol{\psi}}(\widehat{\boldsymbol{\varphi}}^0_{\boldsymbol{\psi}}) \\ \boldsymbol{0} \end{bmatrix}, \quad (7)$$

and the value of $\boldsymbol{s}_{\boldsymbol{\psi}}$ depends on $\boldsymbol{y}^0$. We parameterize this line by $t \in \mathbb{R}$,

$$\boldsymbol{s}(t) = \boldsymbol{s}_{\boldsymbol{\psi}} + t(\boldsymbol{s}^0 - \boldsymbol{s}_{\boldsymbol{\psi}}) = (1-t)\boldsymbol{s}_{\boldsymbol{\psi}};$$

the maximum likelihood estimates $\widehat{\boldsymbol{\varphi}}(\boldsymbol{s})$ in (6) vary with $\boldsymbol{s}(t)$. As t increases, they trace out a curve in the parameter space that passes through the constrained maximum likelihood estimate $\widehat{\boldsymbol{\varphi}}^0_{\boldsymbol{\psi}}$ when $t = 0$ and through the full maximum likelihood estimate $\widehat{\boldsymbol{\varphi}}^0$ when $t = 1$. The mapping of the line $\mathcal{L}^*$ to the maximum likelihood estimates $\widehat{\boldsymbol{\varphi}}(t)$ is indicated in Figure 2.

The saddlepoint approximation constrained to $\mathcal{L}^*$ is simply obtained from (6) as

$$h(t;\boldsymbol{\psi}) = \dot{f}\{\boldsymbol{s}(t);\boldsymbol{\psi}\} = c\exp\left(\ell\left\{\widehat{\boldsymbol{\varphi}}^0_{\boldsymbol{\psi}};\boldsymbol{s}(t)\right\} - \ell[\widehat{\boldsymbol{\varphi}}\{\boldsymbol{s}(t)\};\boldsymbol{s}(t)]\right) \times |\boldsymbol{J}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}[\widehat{\boldsymbol{\varphi}}\{\boldsymbol{s}(t)\};\boldsymbol{s}(t)]\,|^{-1/2}. \quad (8)$$

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

To measure departure from the null hypothesis along this line, we now determine the conditional distribution of $\|\boldsymbol{s}\|$, given the unit vector $\boldsymbol{a} = \boldsymbol{s}/\|\boldsymbol{s}\|$. This conditional distribution is obtained from (6) by a change of variables from $\boldsymbol{s}$ to $(\|\boldsymbol{s}\|, \boldsymbol{a})$. The Jacobian of the transformation—which is on $\mathbb{R}^d$, not $\mathbb{R}^p$, because $\boldsymbol{s}_2$ is fixed on $\mathcal{L}^*$—is proportional to t^{d-1}, as t represents the radial distance, and $\boldsymbol{a}$ the vector of $d-1$ angles. The resulting expression from (8) is

$$t^{d-1}h(t;\boldsymbol{\psi}) \propto t^{d-1}\exp[\ell\{\widehat{\boldsymbol{\varphi}}^0_{\psi};\boldsymbol{s}(t)\} - \ell\{\widehat{\boldsymbol{\varphi}};\boldsymbol{s}(t)\}] \times |J_{\boldsymbol{\varphi}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\varphi}})|^{-1/2}, t \in \mathbb{R}; \quad (9)$$

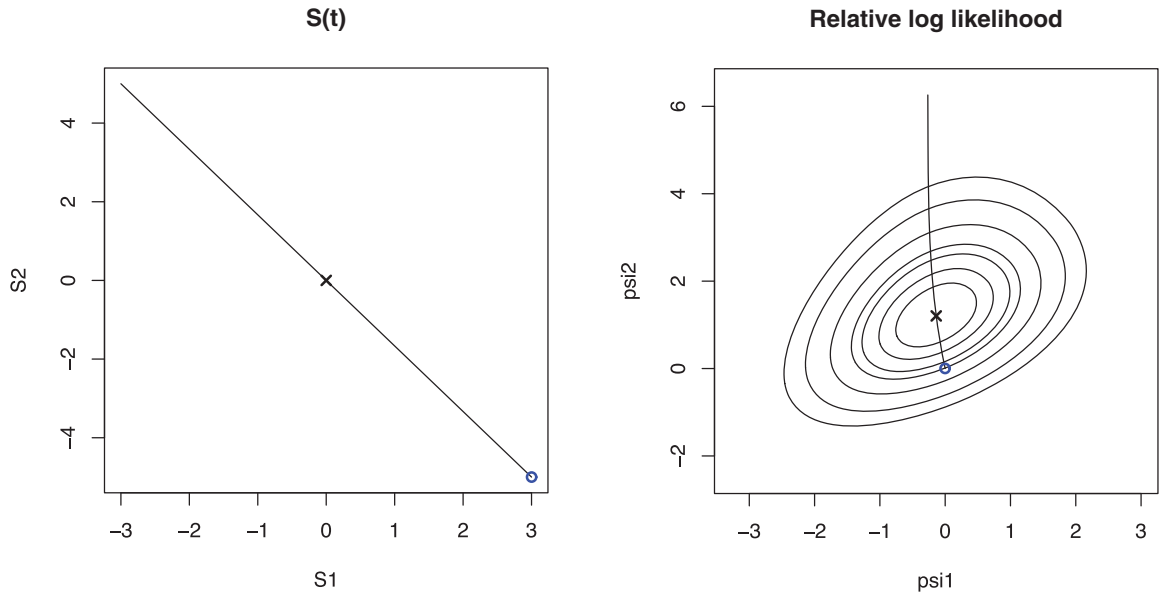


Figure 2. The line $\mathcal{L}^*$ for testing $\boldsymbol{\psi} = \mathbf{0}$, using the data of Table 1. In the plane $\mathcal{L}_0$ on the left, the axes show the two components of $\boldsymbol{s}_1$, as $\boldsymbol{s}_2 = 0$ on $\mathcal{L}_0$. The line $\mathcal{L}^*$ joins the (first two) components of $\boldsymbol{s}_{\boldsymbol{\psi}}$ (○) and of $\boldsymbol{s}^0$ (×); see (7). In the right panel, the observed profile log-likelihood contours are plotted in (ψ_1, ψ_2); the constrained maximum likelihood estimate under H_ψ, and the full maximum likelihood estimate are marked with the same symbols as in the left panel. The curve in the right panel is the mapping of $\mathcal{L}^*$ to the parameter space.

$\widehat{\boldsymbol{\varphi}}$ depends on t through $\boldsymbol{s}$, which is constrained to the line $\mathcal{L}^*$ in $\mathcal{L}^0$.

More detail is given in Cheah, Fraser, and Reid (1994), where the variables $\|\boldsymbol{s}\| \in \mathbb{R}$ and $\boldsymbol{a} \in \mathbb{R}^{d-1}$ are related to parameters ω and $\boldsymbol{v}$, where $(\omega, \boldsymbol{v})$ is a transformation of $\boldsymbol{\psi}$, with ω having the role of a scalar parameter of interest and $\boldsymbol{v}$ that of a $d-1$-dimensional nuisance parameter. The conditional density of $\|\boldsymbol{s}\|$ given $\boldsymbol{a}$ is obtained from a change of variables in (6). The accuracy of (9) is thus inherited from that of the original saddlepoint approximation. This requires smoothness assumptions on the underlying model, including the assumption that the log-likelihood function has a unique maximum in the interior of the parameter space.

The directional test uses $h(t; \boldsymbol{\psi})$ to compute the p-value as the probability that $\boldsymbol{s}(t)$ is as far or farther from $\boldsymbol{s}_{\boldsymbol{\psi}}$ than is the observed value $\mathbf{0}$; this distribution is on the part of $\mathcal{L}^*$ for which $t > 0$. The directed p-value is thus

$$p(\boldsymbol{\psi}) = \frac{\int_1^{t_{\max}} t^{d-1} h(t; \boldsymbol{\psi})\, dt}{\int_0^{t_{\max}} t^{d-1} h(t; \boldsymbol{\psi})\, dt}, \tag{10}$$

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

The upper limit of the integrals in (10) is the largest value of t for which the maximum likelihood estimator corresponding to $\boldsymbol{s}(t)$ exists; for instance, $t_{\max} = 2$ in the example of Figure 1, though $t_{\max}$ may be infinite in some cases. Figure 3 shows the contours of the log-likelihood function $\ell\{\boldsymbol{\varphi}; \boldsymbol{s}(t)\}$ 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 (10) stems from that of the renormalized saddlepoint approximation (6), 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 that 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 (10), for example, in the normal distribution settings of Section 5.1, the approximation seems to be essentially exact.

Directional tests were proposed in Fraser and Massam (1985) as a means of simplifying computations of high-dimensional marginal densities by replacing them with one-dimensional conditional densities, in the context of marginal inference for regression parameters in nonnormal linear models. Skovgaard (1988) proposed directional testing as a means of using the saddlepoint approximation to a sample mean in $\mathbb{R}$, arguing that it is first order equivalent to the likelihood ratio test, but would give a more accurate approximation to the p-value, especially for small p-values, because saddlepoint methods maintain their accuracy into the tails of the distribution. His directional tests also used the length of the score variable; he suggested this for quite general models, although his examples are linear exponential families. Unfortunately, as mentioned in Section 1, approximating the integral that defines the p-value involves as base distribution a χ^2 distribution, rather than a normal, and the accuracy of the resulting approximation to the p-value is poorer than that of numerical integration. Approximations similar to those of Skovgaard (1988) were developed in Cheah, Fraser, and Reid (1994), but the expressions in both articles are cumbersome and can be inaccurate.

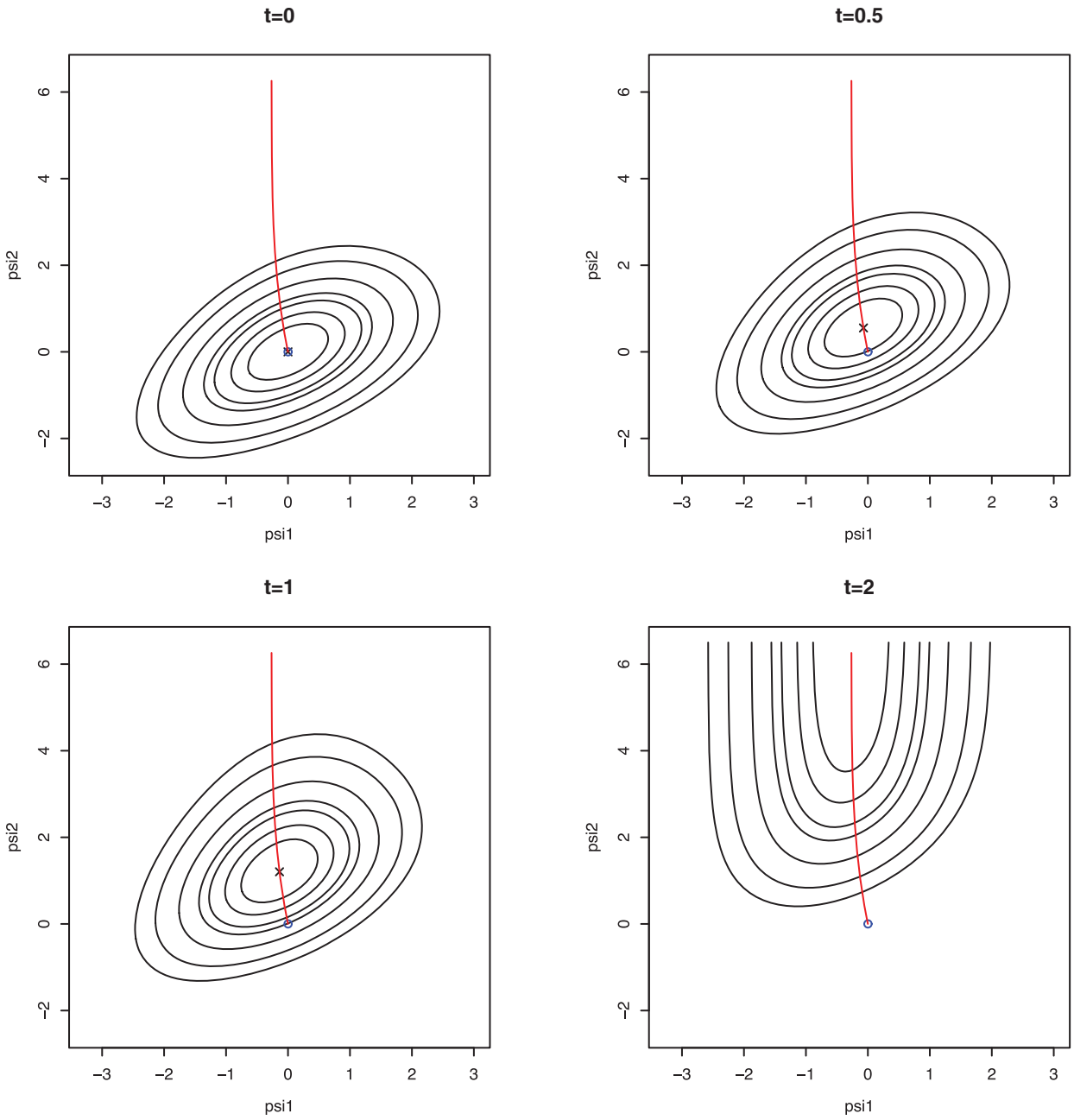


Figure 3. The tilted log-likelihood function, (4), used in (10), and the (red) curve in the parameter space traced by the maximum likelihood estimate as $s(t)$ varies along $\mathcal{L}^*$, for the data and model of Table 1. Each contour plot shows a point on this curve for the value of t indicated above it; the corresponding 2×3 tables are in the right column of Figure 1. The maximum likelihood estimate $\widehat{\boldsymbol{\psi}}\{s(t)\}$ is marked with $\times$, and $\widehat{\boldsymbol{\psi}}\{s(0)\}$ with $\circ$. The lower left plot corresponds to the observed table, and the left panel of Figure 2.

3.3 Skovgaard's Adjusted Likelihood Ratio Statistic

Skovgaard (2001) suggested a direct adjustment to the log-likelihood ratio statistic $w(\boldsymbol{\psi})$, given at (1), as an alternative to directional tests. This was motivated by the arguments used to derive the r^* approximation for inference about a scalar parameter. The adjustment takes the form given in (2), where γ is given in Skovgaard (2001, eq. (10)), and simplified to exponential families in his eq. (13). In our notation, the expression for γ is

$$\gamma = \frac{\{(\boldsymbol{s}-\boldsymbol{s}_{\boldsymbol{\psi}})^\top \boldsymbol{J}^{-1}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{\psi}})(\boldsymbol{s}-\boldsymbol{s}_{\boldsymbol{\psi}})\}^{d/2}}{w^{d/2-1}(\widehat{\boldsymbol{\varphi}}-\widehat{\boldsymbol{\varphi}}_{\boldsymbol{\psi}})^\top(\boldsymbol{s}-\boldsymbol{s}_{\boldsymbol{\psi}})} \left\{ \frac{|\boldsymbol{J}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{\psi}})|}{|\boldsymbol{J}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}(\widehat{\boldsymbol{\varphi}})|} \right\}^{1/2} \quad (11)$$

and for calculating the p-value it would be evaluated at $\boldsymbol{s} = \boldsymbol{0}$, corresponding to $\boldsymbol{y} = \boldsymbol{y}^0$.

Expression (11) adjusts w, at least approximately, by a factor that depends on the ratio of $w(\boldsymbol{\psi})$ to the quadratic form for the score test, and in this sense is closely related to the r^* approximation for scalar parameters. On the other hand, it continues to be an "omnibus" test, in the sense that it attempts to correct the likelihood ratio statistic everywhere in the parameter space. The directional test is designed to capture the alternative suggested by the observed data, although the simulations show that it has good unconditional properties as well.

4. MODELS WITH DISCRETE RESPONSES

4.1 Contingency Tables

The calculations are particularly straightforward for a generic contingency table, as in the example in Section 2. Denote the observed cell frequencies by $\boldsymbol{y}^0 = (y_1, \ldots, y_C)$, 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 $\boldsymbol{X}$ and $\boldsymbol{\theta}$ 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 $\boldsymbol{\mu}(\boldsymbol{\theta}) = \exp(\boldsymbol{X\theta})$, such expressions being understood coordinatewise.

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

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

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

$$\ell^0_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) = \boldsymbol{X}^\top(\boldsymbol{y}^0 - e^{\boldsymbol{X\varphi}}) = \boldsymbol{X}^\top\{\boldsymbol{y}^0 - \boldsymbol{\mu}(\boldsymbol{\varphi})\},$$
$$\boldsymbol{J}_{\boldsymbol{\varphi\varphi}}(\boldsymbol{\varphi}) = \boldsymbol{X}^\top \mathrm{diag}(e^{\boldsymbol{X\varphi}})\boldsymbol{X} = \boldsymbol{X}^\top \mathrm{diag}\{\boldsymbol{\mu}(\boldsymbol{\varphi})\}\boldsymbol{X}.$$

For inference about a component parameter $\boldsymbol{\psi}$ of $\boldsymbol{\varphi}$, the columns of the design matrix are partitioned as $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$, in conformity with $\boldsymbol{\varphi} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$. The hypothesis $H_0 : \boldsymbol{\psi} = \mathbf{0}$ corresponds to the equivalence of the two nested models with linear predictors $\boldsymbol{X}_2\boldsymbol{\lambda}$ and $\boldsymbol{X\varphi}$. The constrained maximum likelihood estimate of $\boldsymbol{\varphi}$ satisfies $\ell_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\varphi}}^0_{\boldsymbol{\psi}}; \boldsymbol{y}^0) = \boldsymbol{X}_2^\top(\boldsymbol{y}^0 - e^{\boldsymbol{X}\widehat{\boldsymbol{\varphi}}^0_{\boldsymbol{\psi}}}) = \mathbf{0}$.

For directional assessment of the null hypothesis, the observed data point $\boldsymbol{s}^0$ equals $\mathbf{0}$ and the expected value $\boldsymbol{s}_{\boldsymbol{\psi}}$ defined in (7) is

$$\boldsymbol{s}_{\boldsymbol{\psi}} = \begin{bmatrix} -\boldsymbol{X}_1^\top(\boldsymbol{y}^0 - e^{\boldsymbol{X}\widehat{\boldsymbol{\varphi}}_{\boldsymbol{\psi}}}) \\ \mathbf{0} \end{bmatrix} = -\boldsymbol{X}^\top\{\boldsymbol{y}^0 - \boldsymbol{\mu}(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{\psi}})\}.$$

The directional p-value is obtained numerically from (10). To determine $t_{\max}$, note that the maximum likelihood estimate satisfies the condition

$$\boldsymbol{X}^\top \boldsymbol{y}^0 = \boldsymbol{X}^\top \boldsymbol{\mu}(\widehat{\boldsymbol{\varphi}}^0) : \tag{12}$$

the observed value of the sufficient statistic equals the expected value under the assumed model. In a contingency table, Equation (12) 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, sec. 9.8.2). When we need to compute $\widehat{\boldsymbol{\varphi}}(t)$, which maximizes $\ell(\boldsymbol{\varphi}; t) = \ell^0(\boldsymbol{\varphi}) + \boldsymbol{\varphi}^\top \boldsymbol{s}(t)$, Equation (12) becomes

$$\boldsymbol{X}^\top\{\widehat{\boldsymbol{\mu}}^0_{\boldsymbol{\psi}} + t(\widehat{\boldsymbol{\mu}}^0 - \widehat{\boldsymbol{\mu}}^0_{\boldsymbol{\psi}})\} = \boldsymbol{X}^\top \widehat{\boldsymbol{\mu}}(t)\,, \tag{13}$$

where $\widehat{\boldsymbol{\mu}}^0 = \boldsymbol{\mu}(\widehat{\boldsymbol{\varphi}}^0)$, $\widehat{\boldsymbol{\mu}}^0_{\boldsymbol{\psi}} = \boldsymbol{\mu}(\widehat{\boldsymbol{\varphi}}^0_{\boldsymbol{\psi}})$, and $\widehat{\boldsymbol{\mu}}(t) = \boldsymbol{\mu}\{\widehat{\boldsymbol{\varphi}}(t)\}$. For any given value of t larger than 1, the maximum likelihood estimate $\widehat{\boldsymbol{\varphi}}(t)$ and corresponding mean parameter $\widehat{\boldsymbol{\mu}}(t)$ can be easily obtained by solving (13) using iteratively reweighted least squares. A value of t is admissible if the corresponding fitted cell frequencies $\widehat{\boldsymbol{\mu}}(t)$ are all nonnegative and the marginal totals implied by (13) are all positive; $t_{\max}$ is the largest such value of t.

Table 2. Sexual enjoyment data (Kolassa and Tanner 1994, sec. 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

Furthermore, if the larger model is saturated, $\boldsymbol{X}$ will be an invertible matrix of dimension $C \times C$. Then (13) simplifies to

$$\widehat{\boldsymbol{\mu}}(t) = \widehat{\boldsymbol{\mu}}^0_{\boldsymbol{\psi}} + t\left(\widehat{\boldsymbol{\mu}}^0 - \widehat{\boldsymbol{\mu}}^0_{\boldsymbol{\psi}}\right),$$

and for the value of t to be admissible each element of $\widehat{\boldsymbol{\mu}}(t)$ must be positive, that is,

$$t < t_{\max} = \min_{i:(\widehat{\boldsymbol{\mu}}^0_{\boldsymbol{\psi}} - \widehat{\boldsymbol{\mu}}^0)_i > 0} \frac{\left(\widehat{\boldsymbol{\mu}}^0_{\boldsymbol{\psi}}\right)_i}{\left(\widehat{\boldsymbol{\mu}}^0_{\boldsymbol{\psi}}\right)_i - (\widehat{\boldsymbol{\mu}}^0)_i}.$$

The directional p-value is obtained from (10) with $h(t; \boldsymbol{\psi})$ given in (8), here equal to

$$h(t; \boldsymbol{\psi}) = \exp\left[\mathbf{1}_C^\top\left\{\widehat{\boldsymbol{\mu}}(t) - \widehat{\boldsymbol{\mu}}^0_{\boldsymbol{\psi}}\right\} - \widehat{\boldsymbol{\mu}}(t)^\top\left\{\log\widehat{\boldsymbol{\mu}}(t) - \log\widehat{\boldsymbol{\mu}}^0_{\boldsymbol{\psi}}\right\}\right] \times |\boldsymbol{X}^\top \mathrm{diag}\{\widehat{\boldsymbol{\mu}}(t)\}\boldsymbol{X}|^{-1/2}.$$

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 $\boldsymbol{\lambda}$ and $\boldsymbol{\psi}$ have dimensions 7 and 9; again we test the null hypothesis $H_0 : \boldsymbol{\psi} = \mathbf{0}$ of independence. The directional p-value (10) is 0.139, while the first-order and Skovgaard's w^* p-values are, respectively, 0.078 and 0.165. Kolassa and Tanner (1994, sec. 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 eight-dimensional parameter $\boldsymbol{\lambda}$ 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 (10) 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

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 (10) or (2) for 4747 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, (10)	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, (10)	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, (10)	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

directional p-values are unconditionally very accurate, as are Skovgaard's (2001) 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 that when equal to zero yield independence of the row and column classifications. In 100,000 simulations with total sample size $n = 150$, 4747 of the simulated tables had a cell margin of zero, in which case neither the directional method nor Skovgaard's (2001) 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's (2001) 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's (2001) large-deviation statistic and the directional test yield essentially identical p-values. This is not the case in general, however, as is seen in Section 5.

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, that is, the density of urine relative to water; pH (ph); osmolarity (mOsm); conductivity (mMho); urea concentration (millimoles per liter); and calcium concentration (millimoles per liter). In the following analysis, we use the $n = 77$ complete observations. A natural starting point for analysis is a logistic regression model with

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

where $\boldsymbol{x}_i$ represents the vector of explanatory variables associated with the ith response y_i. The log-likelihood is of linear exponential form with canonical parameter $\boldsymbol{\varphi} = \boldsymbol{\theta}$, that is,

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

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

The development of the directional p-value is similar to that for contingency tables in Section 4.1. In particular, to compute $\widehat{\boldsymbol{\varphi}}(t)$ we again solve Equation (13) through iterative weighted least squares, but now with $\boldsymbol{\mu}(\boldsymbol{\varphi}) = \exp(\boldsymbol{X}\boldsymbol{\varphi})/\{1 + \exp(\boldsymbol{X}\boldsymbol{\varphi})\}$. In this case, the largest admissible value $t_{\max}$ is the largest value of t for which all fitted probabilities in $\widehat{\boldsymbol{\mu}}(t)$ lie in the interval $[0, 1]$. The function $h(t; \boldsymbol{\psi})$ in (10), given by (8), is then

$$\begin{aligned} h(t; \boldsymbol{\psi}) = \exp\big(\widehat{\boldsymbol{\mu}}(t)^\top \big\{\log \widehat{\boldsymbol{\mu}}_{\boldsymbol{\psi}}^0 - \log \widehat{\boldsymbol{\mu}}(t)\big\} \\ + \{\mathbf{1}_n - \widehat{\boldsymbol{\mu}}(t)\}^\top \big[\log\big(\mathbf{1}_n - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\psi}}^0\big) - \log\{\mathbf{1}_n - \widehat{\boldsymbol{\mu}}(t)\}\big]\big) \\ \times |\boldsymbol{X}^\top \operatorname{diag}[\widehat{\boldsymbol{\mu}}(t)\{1 - \widehat{\boldsymbol{\mu}}(t)\}]\boldsymbol{X}|^{-1/2}. \end{aligned}$$

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, Davison, and Reid (2007, p. 42); there are four nuisance parameters and three interest parameters. The directional p-value (10) 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, sec. 4.2) discussed 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, sec. 10.4), which has a binary

response, presence/absence of bacteria, and measurements on 50 subjects at five times. There are just 24 subjects that are informative for $\boldsymbol{\psi}$, and 108 observations, an average of 4.5 observations per subject. The parameter of interest $\boldsymbol{\psi}$ 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, that is,

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

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 $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_g, \sigma_1^2, \ldots, \sigma_g^2)$ is

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

The full and the constrained maximum likelihood estimates are, respectively,

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

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

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

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),

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

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 Section 6.

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

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

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 can be written as $H_0 : \boldsymbol{\psi} = \mathbf{0}$, with $\psi_j = \varphi_{g+1+j} - \varphi_{g+j}$, $j = 1, \ldots, g-1$. This hypothesis places linear constraints on the canonical parameter $\boldsymbol{\varphi}$, and $K(\boldsymbol{\varphi}) = -\sum_{i=1}^{g} n_i \{2 \log(-2\varphi_{g+i}) + \varphi_i^2 \varphi_{g+i}^{-1}\}/4$. The global and constrained maximum likelihood estimates are

$$\widehat{\boldsymbol{\varphi}}^\top = \left(\frac{\bar{y}_1}{v_1^2}, \ldots, \frac{\bar{y}_g}{v_g^2}, -\frac{1}{2v_1^2}, \ldots, -\frac{1}{2v_g^2}\right), \quad (15)$$

$$\widehat{\boldsymbol{\varphi}}_0^\top = \left(\frac{\bar{y}_1}{\bar{v}^2}, \ldots, \frac{\bar{y}_g}{\bar{v}^2}, -\frac{1}{2\bar{v}^2}, \ldots, -\frac{1}{2\bar{v}^2}\right), \quad (16)$$

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

For the computation of the directional p-value, we need the tilted log-likelihood $\ell(\boldsymbol{\varphi}; s) = \ell^0(\boldsymbol{\varphi}) + \boldsymbol{\varphi}^\top \boldsymbol{s}$, as at (4), where $\boldsymbol{s}^0 = \mathbf{0}$ and

$$\boldsymbol{s}_{\boldsymbol{\psi}} = -\ell^0_{\boldsymbol{\varphi}}(\widehat{\boldsymbol{\varphi}}_0) = \left\{0, \ldots, 0, -n_1\left(v_1^2 - \bar{v}^2\right), \ldots, -n_g\left(v_g^2 - \bar{v}^2\right)\right\}.$$

In this example, the log-likelihood along the line $\boldsymbol{s}(t) = t\boldsymbol{s}^0 + (1-t)\boldsymbol{s}_{\boldsymbol{\psi}} = (1-t)\boldsymbol{s}_{\boldsymbol{\psi}}$ that joins the expected value $\boldsymbol{s}_{\boldsymbol{\psi}}$ and the observed value $\boldsymbol{s}^0$ can be computed explicitly, giving

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

which is maximized at

$$\widehat{\varphi}_i(t) = \frac{\bar{y}_i}{t v_i^2 + (1-t)\bar{v}^2}, \quad \widehat{\varphi}_{g+i}(t) = -\frac{1}{2\{t v_i^2 + (1-t)\bar{v}^2\}}, \quad i = 1, \ldots, g. \quad (17)$$

As expected, $t = 0$ and $t = 1$ give (16) and (15), respectively. Moreover, since $\widehat{\varphi}_{g+i}(t)$ must be negative for all $i = 1, \ldots, g$, we have that

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

$\boldsymbol{s}(t_{\max})$ is the last value of $\boldsymbol{s}$ along the line $\boldsymbol{s}(t)$ that leads to an admissible maximum likelihood estimate (17). The directional p-value is computed from (10), with (8) giving

$$h(t; \boldsymbol{\psi}) \propto \prod_{i=1}^{g} \left\{t v_i^2 + (1-t)\bar{v}^2\right\}^{(n_i - 3)/2}.$$

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

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

Table 4. Data used to illustrate comparison of variances (NIST 2012)

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
$10^5\widehat{\sigma}_i^2$	1.70	2.45	1.42	1.34	5.17	8.80	5.59	1.18	1.54	2.56

NOTE: Sufficient statistics for the gear diameter measurement of $g = 10$ batches each of $n = 10$ observations.

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's (2001) $w^* = r^{*2}$, 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 overcorrect.

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 4. In the left panel, $g = 3$ and $n_i = 5$, giving two interest parameters and four nuisance parameters. As might be guessed from 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 of $\tilde{w}$, although the p-values in individual cases can be different. 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. Inferences based on the likelihood ratio statistic or on Skovgaard's (2001) statistic w^* break down completely, but Bartlett's test and the directional test maintain their level extremely well. In further simulations with the more realistic values $g = 10$ and $n_i = 20$ (not shown), the likelihood ratio test and Skovgaard's statistic are noticeably nonuniform, whereas the directional test and Bartlett's test are essentially exact. In fact the directional test seems to duplicate Bartlett's test for homogeneity of variances, but we have not been able to verify this analytically.

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^{**}.

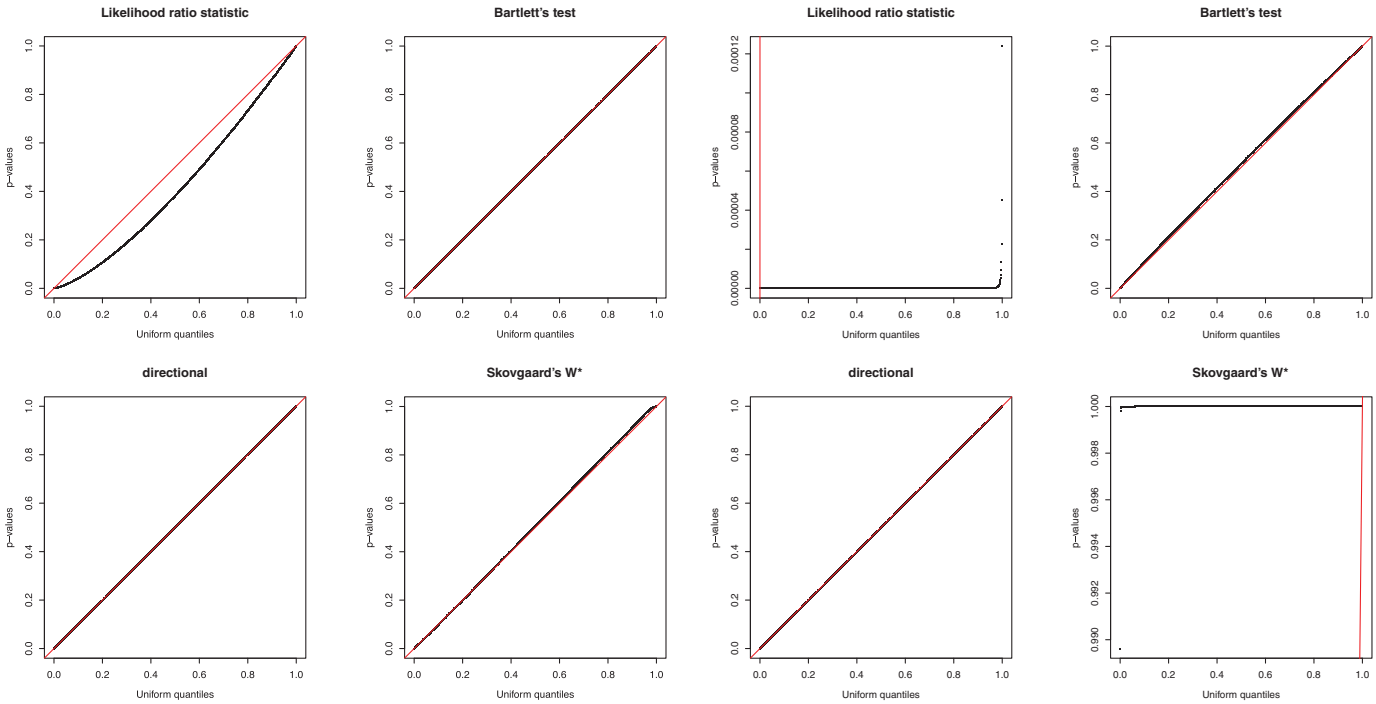


Figure 4. Simulations for testing common variances in $g = 3$ groups with $n_i = 5$ observations per group (left panels), and in $g = 1000$ groups with $n_i = 5$ observations per group (right panels), based on 100,000 replications. We compare the simulated p-values under the null hypothesis to the uniform distribution.

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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_g)$ is

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

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

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

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,

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

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

$$w = 2\sum_{i=1}^{g} n_i \log(\widehat{\theta}_i/\widehat{\theta}^0) = 2\sum_{i=1}^{g} n_i \log(\bar{y}/\bar{y}_i), \tag{18}$$

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

The tilted log-likelihood (4) along the line $s(t)$,

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

is maximized at

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

The line for the directional test goes through $\boldsymbol{s}^0$ and $\boldsymbol{s}_{\boldsymbol{\psi}} = \{-n_1(\bar{y}_1 - \bar{y}), \ldots, -n_g(\bar{y}_g - \bar{y})\}$, where $\widehat{\boldsymbol{\varphi}}(0) = -\widehat{\boldsymbol{\theta}}^0$, and since $\widehat{\varphi}_i(t)$ 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 (10) uses $h(t; \boldsymbol{\psi})$ from (8), resulting in

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

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

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

$$\gamma = \left\{\sum_{i=1}^{g} n_i \left(\frac{\bar{y}_i - \bar{y}}{\bar{y}}\right)^2\right\}^{(g-1)/2} \left(\prod_{i=1}^{g} \frac{\bar{y}}{\bar{y}_i}\right) \Bigg/ \left\{w^{(g-1)/2} \sum_{i=1}^{g} \frac{n_i(\bar{y}_i - \bar{y})^2}{\bar{y}\bar{y}_i}\right\}. \tag{19}$$

We illustrate these calculations by testing the equality of the mean times between failures of the air-conditioning equipment in 10 Boeing 720 aircraft (Proschan 1963; Cox and Snell 1981). The first-order p-value based on (18) equals 0.0198, the directional p-value (10) equals 0.0227, and Skovgaard's modified likelihood ratio statistic (2) equals 0.0274.

Table 5 summarizes two simulation studies; one using the same sample sizes as in the example, and one using $g = 1000$ groups of size 5, which has 999 parameters of interest, but just one nuisance parameter. The results confirm the very accurate behavior of the directional approach, while showing a substantial worsening of the performance of both the likelihood ratio statistic and, more remarkably, Skovgaard's statistic in the second setting.

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).

Let $y_1, \ldots, y_n$ be a sample of independent random vectors from a multivariate normal $N_q(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, where the mean $\boldsymbol{\mu}$ and

Table 5. Simulated empirical distribution (%) of p-values for testing equality of exponential rates, based on 100,000 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, (18)	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^*, (19)	0.9	2.4	4.7	9.6	23.9	48.2	73.1	88.7	94.2	96.9	98.6
Directional, (10)	1.0	2.6	5.0	10.2	25.0	49.9	74.9	90.0	95.0	97.5	99.0
Likelihood ratio, (18)	5.8	11.3	18.4	29.7	52.6	76.9	91.8	97.7	99.1	99.6	99.9
Skovgaard's w^*, (19)	0.0	0.0	0.0	0.0	0.0	0.2	1.2	4.5	8.9	14.6	23.8
Directional, (10)	1.1	2.6	5.0	10.1	25.3	50.3	75.3	90.1	95.1	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

NOTE: 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 $g = 1000$ groups and sample sizes $n_i = 5$.

the concentration matrix $\boldsymbol{\Lambda}$ are unknown and arbitrary apart from the restriction that $\boldsymbol{\Lambda}$ is symmetric and positive definite. Let $\boldsymbol{y}$ denote the $n \times q$ matrix with ith row vector $\boldsymbol{y}_i^\top$. Then the log-likelihood for $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is

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

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

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

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

$$w = -n \log(|\widehat{\boldsymbol{\Lambda}}^{-1} \widehat{\boldsymbol{\Lambda}}_0|) , \tag{20}$$

follows the χ^2_d distribution.

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

$$\ell(\boldsymbol{\varphi}; y) = \frac{n}{2} \log |\boldsymbol{\Lambda}| - \frac{1}{2} \mathrm{tr}(\boldsymbol{\Lambda} \boldsymbol{y}^\top \boldsymbol{y}) + \mathbf{1}_n^\top \boldsymbol{y} \boldsymbol{\xi} - \frac{n}{2} \boldsymbol{\xi}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{\xi}. \tag{21}$$

The expected value $\boldsymbol{s}_{\boldsymbol{\psi}}$ defined in (7) is $\boldsymbol{s}_{\boldsymbol{\psi}} = -\{\ell_{\boldsymbol{\xi}}(\widehat{\boldsymbol{\varphi}}_0), \ell_{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\varphi}}_0)\} = \{\mathbf{0}, n(\widehat{\boldsymbol{\Lambda}}^{-1} - \widehat{\boldsymbol{\Lambda}}_0^{-1})/2\}$. The tilted log-likelihood (4) along the line $\boldsymbol{s}(t) = (1-t)\boldsymbol{s}_{\boldsymbol{\psi}}$ can be obtained using (21). The maximization is straightforward in the $\boldsymbol{\theta}$ parameterization and yields $\widehat{\boldsymbol{\theta}}(t) = \{\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Lambda}}(t)\}$, with $\widehat{\boldsymbol{\Lambda}}(t)^{-1} = t\widehat{\boldsymbol{\Lambda}}^{-1} + (1-t)\widehat{\boldsymbol{\Lambda}}_0^{-1}$. The last value of $\boldsymbol{s}$ along the line $\boldsymbol{s}(t)$, $\boldsymbol{s}(t_{\max})$, is the largest value such that $\widehat{\boldsymbol{\Lambda}}(t)$ is positive definite, and this can easily be found numerically.

The directional p-value (10) uses $h(t; \boldsymbol{\psi})$ from (8), and since $|\boldsymbol{J}_{\boldsymbol{\varphi}\boldsymbol{\varphi}}[\widehat{\boldsymbol{\varphi}}\{\boldsymbol{s}(t)\}; \boldsymbol{s}(t)]|^{-1/2} = |\widehat{\boldsymbol{\Lambda}}(t)|^{(q+2)/2}$ we find that

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

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

$$\gamma = \left[\frac{1}{2} \left\{ \mathrm{tr}(\widehat{\boldsymbol{\Lambda}}^{-1} \widehat{\boldsymbol{\Lambda}}_0 \widehat{\boldsymbol{\Lambda}}^{-1} \widehat{\boldsymbol{\Lambda}}_0) - q \right\} \right]^{d/2} |\widehat{\boldsymbol{\Lambda}}^{-1} \widehat{\boldsymbol{\Lambda}}_0|^{-(q+2)/2} \Big/ \frac{1}{2} \left\{ \mathrm{tr}(\widehat{\boldsymbol{\Lambda}} \widehat{\boldsymbol{\Lambda}}_0^{-1}) - q \right\} \left(- \log |\widehat{\boldsymbol{\Lambda}}^{-1} \widehat{\boldsymbol{\Lambda}}_0| \right)^{d/2-1} . \tag{22}$$

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 10 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, that is, we test that all off-diagonal elements of $\boldsymbol{\Lambda}$ are zero, except those closest to the diagonal. In the saturated model $\boldsymbol{\Lambda}$ 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 finding is 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 covariances in the multivariate normal models is sometimes based on the Wishart marginal distribution of the sample covariance matrix, which is free of the nuisance parameters $\boldsymbol{\mu}$; 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 (21). The 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^*.

Table 6. Simulated empirical distribution (%) of p-values for testing first-order Markov dependence with $n = 60$, based on 100, 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, (20)	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^*, (22)	0.7	1.8	3.6	7.4	19.6	42.2	67.8	85.2	91.9	95.5	98.0
Directional, (10)	1.1	2.6	5.0	10.1	24.8	49.8	74.9	89.9	94.9	97.4	99.0
Likelihood ratio, (20)	91.2	95.4	97.5	98.9	99.8	100	100	100	100	100	100
Skovgaard's w^*, (22)	0.0	0.0	0.0	0.2	1.1	4.7	14.8	31.5	44.0	55.6	68.5
Directional, (10)	1.0	2.5	5.0	10.1	25.2	50.2	75.1	90.1	95.0	97.5	99.0
Likelihood ratio, (20)	100	100	100	100	100	100	100	100	100	100	100
Skovgaard's w^*, (22)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.7	1.6	2.9	5.6
Directional, (10)	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

NOTE: 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.

6. DISCUSSION

We have presented the formulas for the conditional density $h(t;\boldsymbol{\psi})$ on the line, (8), and associated p-value, (10), 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 (6). Ongoing work derives such a reference distribution as a marginal density for a derived variable, and shows that the saddlepoint approximation to this density has a form similar to (6), but with an additional adjustment for curvature. The construction of a directional test from this marginal density proceeds in the same way as in the linear case treated here, and Fraser and Reid (2014) illustrate this on a number of examples.

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 $\boldsymbol{\varphi}(\boldsymbol{\theta})$ 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, Reid, and Wu (1999) and Brazzale, Davison, and Reid (2007, chap. 8). The r^*-type approximations provide inference for scalar parameters of interest only. Application of the directional method for inference about vector parameters in tangent exponential models seems relatively straightforward, but as yet we have little experience with concrete applications.

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

$$\tilde{w}(\boldsymbol{\psi}) = w(\boldsymbol{\psi})/[\mathrm{E}_{\boldsymbol{\theta}}\{w(\boldsymbol{\psi})\}/d] \tag{23}$$

follows a χ^2_d distribution with relative error $O(n^{-2})$. Skovgaard (2001) noted that the accuracy of the χ^2_d approximation to (23) 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 Section 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.

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 nonstandard aspect is the determination of $t_{\max}$ in (10). For contingency tables, as discussed in Section 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 Section 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 Section 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. Other Metropolis–Hastings algorithms for conditional simulation can also be implemented (Forster, McDonald, and Smith 1996; Smith, Forster, and McDonald 1996; Diaconis and Sturmfels 1998; Forster, McDonald, and Smith 2003), but ensuring irreducibility of the resulting chain is again not straightforward in general, and so far as we know no general code is available for this. Caffo and Booth (2003) gave a helpful overview of Monte Carlo methods for log-linear models.

Naive bootstrap simulation from the fitted model would be expected to give lower theoretical accuracy than the approach described above, as it is unconditional, rather than conditioned on the sufficient statistics. Although precision can be improved by nested simulation (Davison and Hinkley 1997, sec. 4.5), the computational burden would then greatly increase. A more accurate unconditional approach due to DiCiccio and Young (2008) 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 consideration. Unlike the computational methods mentioned above, it has the added advantage that the same paradigm holds for both discrete and continuous models.

[Received March 2013. Revised August 2013.]

REFERENCES

Agresti, A. (2002), *Categorical Data Analysis* (2nd ed.), New York: Wiley. [307]

Andrews, D. F., and Herzberg, A. M. (1985), *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, New York: Springer. [308]

Barndorff-Nielsen, O. E., and Cox, D. R. (1979), "Edgeworth and Saddle-Point Approximations With Statistical Applications" (with discussion), *Journal of the Royal Statistical Society,* Series B, 41, 279–312. [304]

——— (1994), *Inference and Asymptotics*, London: Chapman and Hall. [309]

Bartlett, M. S. (1937), "Properties of Sufficiency and Statistical Tests," *Proceedings of the Royal Society London A*, 160, 268–282. [309,313]

Birch, M. (1963), "Maximum Likelihood in Three-Way Contingency Tables," *Journal of the Royal Statistical Society,* Series B, 25, 220–223. [307]

Brazzale, A. R., and Davison, A. C. (2008), "Accurate Parametric Inference for Small Samples," *Statistical Science*, 23, 465–484. [302,308]

Brazzale, A. R., Davison, A. C., and Reid, N. (2007), *Applied Asymptotics*, Cambridge: Cambridge University Press. [302,308,313]

Butler, R. W. (2007), *Saddlepoint Approximations with Applications*, Cambridge: Cambridge University Press. [304,305]
Caffo, B. (2006), "Exact Hypothesis Test for Log-Linear Models With `exactLoglinTest`," *Journal of Statistical Software*, 17, 1–17. [313]
Caffo, B., and Booth, J. (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. [313]
Cheah, P., Fraser, D. A. S., and Reid, N. (1994), "Multiparameter Testing in Exponential Models: Third Order Approximations From Likelihood," *Biometrika*, 81, 271–278. [302,305]
Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall. [303,305]
Cox, D. R., and Snell, E. J. (1981), *Applied Statistics: Principles and Examples*, London: Chapman and Hall. [311]
Daniels, H. E. (1958), Discussion of "The Regression Analysis of Binary Sequences," by D. R. Cox, *Journal of the Royal Statistical Society,* Series B, 20, 236–238. [304]
Davison, A. C. (1988), "Approximate Conditional Inference in Generalized Linear Models," *Journal of the Royal Statistical Society,* Series B, 50, 445–461. [304]
Davison, A. C., Fraser, D. A. S., and Reid, N. (2006), "Improved Likelihood Inference for Discrete Data," *Journal of the Royal Statistical Society,* Series B, 68, 495–508. [302]
Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge: Cambridge University Press. [313]
Diaconis, P., and Sturmfels, B. (1998), "Algebraic Algorithms for Sampling From Conditional Distributions," *The Annals of Statistics*, 26, 363–397. [313]
DiCiccio, T. J., and Young, G. A. (2008), "Conditional Properties of Unconditional Parametric Bootstrap Procedures for Inference in Exponential Families," *Biometrika*, 95, 747–758. [313]
Everitt, B. S. (1992), *The Analysis of Contingency Tables* (2nd ed.), London: Chapman and Hall. [303]
Fienberg, S. E., and Rinaldo, A. (2012), "Maximum Likelihood Estimation in Log-Linear Models," *The Annals of Statistics*, 40, 996–1023. [313]
Forster, J. J., McDonald, J. W., and Smith, P. W. F. (1996), "Monte Carlo Exact Conditional Tests for Log-Linear and Logistic Models," *Journal of the Royal Statistical Society,* Series B, 58, 445–453. [313]
——— (2003), "Markov Chain Monte Carlo Exact Inference for Binomial and Multinomial Regression Models," *Statistics and Computing*, 13, 169–177. [313]
Fraser, D. A. S., and Massam, H. (1985), "Conical Tests: Observed Levels of Significance and Confidence Regions," *Statistics Hefte*, 26, 1–17. [302,305]
Fraser, D. A. S., and Reid, N. (1995), "Ancillaries and Third Order Significance," *Utilitas Mathematica*, 47, 33–53. [313]
——— (2006), "Assessing a Vector Parameter," *Student*, 5, 247–256. [303]
——— (2014), Assessing a Parameter of Interest: Bootstrap and Higher Order Methods," submitted. [304,313]
Fraser, D. A. S., Reid, N., and Wu, J. (1999), "A Simple General Formula for Tail Probabilities for Frequentist and Bayesian Inference," *Biometrika*, 86, 249–264. [313]
Kenward, M. G. (1987), "A Method for Comparing Profiles of Repeated Measurements," *Journal of the Royal Statistical Society,* Series C, 36, 296–308. [312]
Kolassa, J. E., and Tanner, M. A. (1994), "Approximate Conditional Inference in Exponential Families via the Gibbs Sampler," *Journal of the American Statistical Association*, 89, 697–702. [303,307]
Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Oxford University Press. [311,312]
Lawley, D. N. (1956), "A General Method for Approximating to the Distribution of the Likelihood Ratio Criteria," *Biometrika*, 43, 295–303. [313]
Lehmann, E. L., and Romano, J. P. (2005), *Testing Statistical Hypotheses*, New York: Springer. [304]
McCullagh, P., and Cox, D. R. (1986), "Invariants and Likelihood Ratio Statistics," *Annals of Statistics*, 14, 1419–1430. [313]
McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall. [302]
Mehta, C. (1991), "StatXact: A Statistical Package for Exact Nonparametric Inference," *The American Statistician*, 45, 74–75. [313]
NIST (2012), "NIST/SEMATECH e-Handbook of Statistical Methods," available at *http://www.itl.nist.gov/div898/handbook/eda/section3/eda357.htm* (accessed October 19, 2012). [310]
Proschan, F. (1963), "Theoretical Explanation of Observed Decreasing Failure Rate," *Technometrics*, 5, 375–383. [311]
R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [313]
Skovgaard, I. M. (1988), "Saddlepoint Expansions for Directional Test Probabilities," *Journal of the Royal Statistical Society,* Series B, 50, 3–32. [302,305]
——— (2001), "Likelihood Asymptotics," *Scandinavian Journal of Statistics*, 28, 3–32. [302,303,306,308,309,310,311,312,313]
Smith, P. W. F., Forster, J. J., and McDonald, J. W. (1996), "Monte Carlo Exact Tests for Square Contingency Tables," *Journal of the Royal Statistical Society,* Series A, 159, 309–321. [313]
Venables, W., and Ripley, B. (2002), *Modern Applied Statistics With S* (4th ed.), New York: Springer. [308]
Wood, A. T. A., Booth, J. G., and Butler, R. W. (1993), "Saddlepoint Approximation to the CDF of Some Statistics With Non-Normal Limit Distributions," *Journal of the American Statistical Association*, 88, 680–686. [302]