J. R. Statist. Soc. B (2006)
68, *Part* 2, *pp.* 281–303

Variance stabilization for a scalar parameter

Thomas J. DiCiccio,

Cornell University, Ithaca, USA

Anna Clara Monti

University of Sannio, Benevento, Italy

and G. Alastair Young

Imperial College London, UK

[Received September 2005. Revised November 2005]

Summary. We present a variance stabilizing transformation for inference about a scalar parameter that is estimated by a function of a multivariate *M*-estimator. The transformation proposed is automatic, computationally simple and can be applied quite generally. Though it is based on an intuitive notion and entirely empirical, the transformation is shown to have an appropriate justification in providing variance stabilization when viewed from both parametric and nonparametric perspectives. Further, the transformation repairs deficiencies of existing methods for variance stabilization. The transformation proposed is illustrated in a range of examples, and its effectiveness to yield confidence limits having low coverage error is demonstrated in a numerical example.

Keywords: Asymptotic expansion; Bootstrap inference; Box–Cox transformation; Confidence limit; Coverage accuracy; Least favourable family; *M*-estimator; Nonparametric likelihood; Profile likelihood; Studentized statistic; Variance parameter plot

1. Introduction

This paper concerns variance stabilization for inference about a scalar parameter that is estimated by a function of a multivariate *M*-estimator; to fix the notation, say that a scalar $\gamma = g(\theta)$ is estimated by $\hat{\gamma} = g(\hat{\theta})$, where θ is a q-dimensional parameter and its estimator $\hat{\theta}$ is obtained by *M*-estimation based on independent observations $X_1, \ldots, X_n$. The principal focus is on nonparametric inference, with a view towards constructing bootstrap confidence limits; however, parametric inference is also considered. A concise description of bootstrap confidence limits has been given by Davison and Hinkley (1997), section 5.2. Two popular methods are the basic percentile method and the bootstrap t method. The basic percentile method is motivated by the assumption that $n^{1/2}(\hat{\gamma} - \gamma)$ is pivotal. The motivating assumption for the bootstrap t method is that $n^{1/2}(\hat{\gamma} - \gamma)/\hat{\sigma}$ is pivotal, where $\hat{\sigma}^2$ is some $n^{1/2}$-consistent estimator of σ^2, the asymptotic variance of $n^{1/2}(\hat{\gamma} - \gamma)$; typically, $\hat{\sigma}^2$ is derived by using the delta method or the bootstrap. There is considerable practical evidence that the basic percentile bootstrap method and the bootstrap t method are likely to be most accurate when applied in terms of a reparameterization $\phi = h(\gamma)$ that is essentially a location parameter, which can be achieved, at least approximately, by a

Address for correspondence: G. Alastair Young, Department of Mathematics, Huxley Building, Imperial College London, 180 Queen's Gate, London, SW7 2AZ, UK.
E-mail: alastair.young@imperial.ac.uk

 1369–7412/06/68281

variance stabilizing transformation; see, for example, Tibshirani (1988), Efron and Tibshirani (1993), section 12.6, Canty *et al.* (1996) and Davison and Hinkley (1997), section 5.7.

A main device for identifying variance stabilizing transformations is the variance parameter plot, which is a scatterplot of the points $(\hat{\gamma}_b^*, \hat{\sigma}_b^{*2})$, $b = 1, \ldots, B$. These points are B-versions of $(\hat{\gamma}, \hat{\sigma}^2)$ based on B bootstrap samples drawn from the fitted model; in the nonparametric case, the fitted model is the empirical distribution function. Tibshirani (1988) recommended drawing a smooth curve through these points to obtain a variance function $\hat{v}(\gamma)$ and then applying the standard transformation for variance stabilization:

$$h(\gamma) = \int_{\hat{\gamma}}^{\gamma} \hat{v}(t)^{-1/2}\, \mathrm{d}t.$$

A drawback of this approach is that $h(\gamma)$ does not have a closed form expression; typically, numerical integration is necessary to compute the variance-stabilized parameterization. Davison and Hinkley (1997), sections 3.9 and 5.2, recommended the variance parameter plot as a diagnostic tool; a convenient transformation is deemed variance stabilizing if no systematic trend is evident in its variance parameter plot. They provided examples demonstrating that, in many situations, very simple transformations, such as logarithms and square roots, provide an adequate degree of variance stabilization. The drawback of their approach is that it does not automatically identify a suitable transformation. In each case, a scatterplot must be examined and some judgment must be used to recommend possible transformations.

The primary goal of the present work is to provide an automatic and general method to identify convenient variance stabilizing transformations in both parametric and nonparametric problems. The method that is proposed here is based on an asymptotic quantity $\hat{k}$ of order $O_p(1)$ such that

$$\hat{k} = \frac{n}{\hat{\sigma}^3} E\{(\hat{\gamma}^* - \hat{\gamma})(\hat{\sigma}^{*2} - \hat{\sigma}^2)\}$$

to error of order $O(n^{-1/2})$ given the sample values, where the expectation is taken with respect to the distribution of the bootstrap random variable $(\hat{\gamma}^*, \hat{\sigma}^{*2})$. Any parameterization for which the variance parameter plot shows no systematic linear trend would have $\hat{k}$ near 0. The method proposed is to choose, from a suitable class of transformations, a reparameterization for which $\hat{k} = 0$. Motivated by the simple transformations that Davison and Hinkley (1997) found to be effective, the class of Box–Cox transformations is used here. However, other classes of transformations could in principle be considered.

An advantageous property of the method proposed is that it requires no bootstrap resampling. If a resampling technique is to be used for inference about γ, then preparing a variance parameter plot is unlikely to demand much additional computation. However, if a nonresampling method, such as the large sample normal approximation to the Studentized statistic $n^{1/2}(\hat{\gamma} - \gamma)/\hat{\sigma}$, is to be used, then the resampling that is required for the variance parameter plot could increase the computational burden unacceptably. Davison and Hinkley (1997), section 5.2, remarked that variance stabilizing transformations can be effective for improving the standard normal approximation to the distribution of the Studentized statistic. Furthermore, saddlepoint approximations allow bootstrap inference without resampling; see Davison and Hinkley (1988). For example, Daniels and Young (1991) and DiCiccio *et al.* (1994) considered saddlepoint approximations to bootstrap distributions of Studentized statistics and, in this context, the techniques that are developed here are especially convenient for choosing an appropriate parameterization in which to express the Studentized statistic.

Another troublesome aspect of the variance parameter plot is that it depends exclusively on the fitted model; no family of models that is indexed by the interest parameter γ is explicitly

considered. Consequently, it is difficult to understand precisely the sense in which transformations that are derived from variance parameter plots are variance stabilizing. This problem is overcome in an alternative approach to variance stabilizing transformations that are based on least favourable families.

Frequentist inference about γ in the presence of nuisance parameters may be achieved by considering a data-dependent subclass of models that is indexed by γ known as a least favourable family. In the nonparametric context, a least favourable family is a class of multinomial distributions that is determined by a vector of probabilities defined on the data points $X_1, \ldots, X_n$. In parametric models where maximum likelihood estimation is used, a least favourable family is given by the fitted model under the constrained maximum likelihood estimator of θ for a given value of γ. Let $\tilde{\sigma}^2(\gamma)$ be the asymptotic variance function for inference about the interest parameter under sampling from the least favourable family corresponding to the specified value of γ. A key property of the least favourable family is that $\tilde{\sigma}^2(\hat{\gamma})$ is an $n^{1/2}$-consistent estimator of σ^2. Assume that the estimator $\hat{\sigma}^2$ satisfies

$$\hat{\sigma}^2 = \tilde{\sigma}^2(\hat{\gamma}) + O(n^{-1})$$

given the sample values; typically, $\hat{\sigma}^2 = \tilde{\sigma}^2(\hat{\gamma})$. For nonparametric inference, DiCiccio and Tibshirani (1987) and Hall and Presnell (1999) suggested that variance stabilizing transformations be constructed by using this variance function in the standard way. Unfortunately, as in the case of variance parameter plots, this approach typically yields transformations that have no closed form expression, so numerical integration is again required.

Another major aim of the present paper is to establish a formal connection between the two approaches to variance stabilization. In particular, it is shown that

$$\hat{k} = \frac{1}{\hat{\sigma}} \frac{\mathrm{d}\tilde{\sigma}^2(\gamma)}{\mathrm{d}\gamma}\bigg|_{\gamma=\hat{\gamma}} \tag{1.1}$$

to error of order $O(n^{-1/2})$ given the sample values. In many important cases, identity (1.1) holds exactly. It follows from expression (1.1) that, given the sample values,

$$\frac{\tilde{\sigma}^2(\gamma)}{\tilde{\sigma}^2(\hat{\gamma})} = 1 - n^{-1/2} t \hat{k} + O(n^{-1}), \tag{1.2}$$

for values of γ such that $\gamma - \hat{\gamma}$ is of order $O(n^{-1/2})$, where $t = n^{1/2}(\hat{\gamma} - \gamma)/\hat{\sigma}$ is the usual Studentized statistic. Expression (1.2) shows that

$$\tilde{\sigma}^2(\gamma) = \tilde{\sigma}^2(\hat{\gamma}) + O(n^{-1/2})$$

in general; however, if the parameter γ has $\hat{k} = 0$, then

$$\tilde{\sigma}^2(\gamma) = \tilde{\sigma}^2(\hat{\gamma}) + O(n^{-1}).$$

Thus, identity (1.1) establishes that the method of variance stabilization based on variance parameter plots offers a considerable degree of variance stabilization along the least favourable family locally near $\hat{\gamma}$, as does the automatic method of variance stabilization that is proposed here.

The condition $\hat{k} = 0$ does not, in general, imply exact variance stabilization. In practice, even when a reparameterization is used for which the second term on the right-hand side of equation (1.2) vanishes, the effect of higher order terms, particularly the quadratic term, can still be apparent. Variance parameter plots for the transformed parameter often show a parabolic trend; see, for example, the right-hand panels of Fig. 12.2 of Efron and Tibshirani (1993), page

166, and Fig. 5.1 of Davison and Hinkley (1997), page 201. None-the-less, the extent of fluctuation in the variance across the relevant range of the interest parameter is typically much less for the transformed parameter than it is in the original parameterization.

Expansion (1.2) provides another interpretation of $\hat{k}$. Suppose that $\phi = h(\gamma)$ is some reparameterization, and let $\hat{\phi} = h(\hat{\gamma})$. In obvious notation,

$$\begin{aligned}\frac{\tilde{\sigma}_\phi^2(\phi)}{\tilde{\sigma}_\phi^2(\hat{\phi})} &= 1 - \frac{(\hat{\phi}-\phi)}{\hat{\sigma}_\phi}\hat{k}_\phi + O(n^{-1}) \\ &= 1 - \frac{(\hat{\gamma}-\gamma)}{\hat{\sigma}_\gamma}\hat{k}_\phi + O(n^{-1}),\end{aligned}$$

since

$$(\hat{\phi}-\phi)/\hat{\sigma}_\phi = (\hat{\gamma}-\gamma)/\hat{\sigma}_\gamma + O(n^{-1}).$$

It follows that

$$\hat{k}_\phi = \frac{n^{1/2}}{c}\left[\frac{\tilde{\sigma}_\phi^2\{h(\hat{\gamma}+n^{-1/2}c\hat{\sigma}_\gamma)\} - \tilde{\sigma}_\phi^2\{h(\hat{\gamma})\}}{\tilde{\sigma}_\phi^2\{h(\hat{\gamma})\}}\right] + O(n^{-1/2}),$$

for any constant c, since $\tilde{\sigma}_\phi^2(\hat{\phi}) = \tilde{\sigma}_\phi^2\{h(\hat{\gamma})\} = \hat{\sigma}_\phi^2 + O(n^{-1})$. Consequently, $n^{-1/2}c\hat{k}_\phi$ represents, to error of order $O(n^{-1})$, the relative change in variance on the ϕ-scale between the values $\hat{\gamma}+n^{-1/2}c\hat{\sigma}_\gamma$ and $\hat{\gamma}$ of the interest parameter γ. This observation implies that the quantity $\hat{k}$ can be used directly as a guide to distinguish between competing parameterizations in terms of their ability to stabilize variance. If $\phi_1 = h_1(\gamma)$ and $\phi_2 = h_2(\gamma)$ are two reparameterizations having $|\hat{k}_{\phi_1}| < |\hat{k}_{\phi_2}|$, then $\tilde{\sigma}_{\phi_1}^2$ is likely to be relatively less variable locally near $\hat{\phi}_1$ than $\tilde{\sigma}_{\phi_2}^2$ is near $\hat{\phi}_2$.

The fundamentals of our proposed method for variance stabilization are described in Section 2, where examples are given which compare the results of our approach with results from other variance stabilization techniques. In Section 3, the particular case of maximum likelihood estimation in parametric models is considered; the formal sense in which variance stabilization is achieved is elucidated and a numerical study is used to demonstrate the effectiveness of the proposed method for yielding confidence limits having low coverage error. These ideas and results are extended to nonparametric problems in Section 4.

2. A variance stabilizing transformation

2.1. The inference problem

Let X be a random vector, and let $\theta = (\theta^1, \ldots, \theta^q)'$ be an M-estimand defined by

$$E\{\psi(X,\theta)\} = 0,$$

where $\psi(X,\theta) = (\psi_1(X,\theta), \ldots, \psi_q(X,\theta))'$ is a vector of estimating functions. Given a random sample $X_1, \ldots, X_n$ from the same distribution as that of X, the M-estimator $\hat{\theta}$ of θ satisfies the equation

$$\sum_{i=1}^{n} \psi(X_i, \hat{\theta}) = 0.$$

Now suppose that the scalar parameter of interest is $\gamma = g(\theta)$, where $g(\theta)$ is a smooth function, and let $\hat{\gamma} = g(\hat{\theta})$. Under regularity conditions (Huber (1981), page 132), $n^{1/2}(\hat{\theta}-\theta)$ is normally distributed asymptotically, so the asymptotic distribution of $n^{1/2}(\hat{\gamma}-\gamma)$ is $N(0, \sigma^2)$.

The formulae for σ^2 and k require some additional notation. Let

$$g_a(\theta) = \partial g(\theta)/\partial\theta^a,$$
$$g_{ab}(\theta) = \partial^2 g(\theta)/\partial\theta^a\,\partial\theta^b,$$
$$\psi_{a/b}(X,\theta) = \partial\psi_a(X,\theta)/\partial\theta^b,$$
$$\psi_{a/bc}(X,\theta) = \partial^2\psi_a(X,\theta)/\partial\theta^b\,\partial\theta^c,$$

and so forth $(a, b = 1, \ldots, q)$, and let

$$A_{a/b} = E\{\psi_{a/b}(X,\theta)\},$$
$$\Sigma_{ab} = E\{\psi_a(X,\theta)\,\psi_b(X,\theta)\}.$$

Furthermore, define $q \times q$ matrices $A = (A_{a/b})$, $A^{-1} = (A^{a/b})$, $\Sigma = (\Sigma_{ab})$ and $V = (V^{ab}) = A^{-1}\Sigma(A^{-1})'$. Note that V is the asymptotic covariance matrix of $n^{1/2}(\hat{\theta} - \theta)$. Then

$$\sigma^2 = g'_{(1)} V g_{(1)},$$

where $g_{(1)} = (g_1, \ldots, g_q)'$.

Estimates of the preceding quantities can be obtained by replacing expectations with their corresponding sample averages and substituting $\hat{\theta}$ for θ; a circumflex is used to denote such estimates. Thus, $\hat{A} = (\hat{A}_{a/b})$ and $\hat{\Sigma} = (\hat{\Sigma}_{ab})$, where

$$\hat{A}_{a/b} = n^{-1}\sum \psi_{a/b}(X_i, \hat{\theta})$$

and

$$\hat{\Sigma}_{ab} = n^{-1}\sum \psi_a(X_i, \hat{\theta})\,\psi_b(X_i, \hat{\theta});$$

$\hat{g}_{(1)} = (\hat{g}_1, \ldots, \hat{g}_q)'$, where $\hat{g}_a = g_a(\hat{\theta})$. In particular, $\hat{\sigma}^2 = \hat{g}'_{(1)}\hat{V}\hat{g}_{(1)}$, where $\hat{V} = (\hat{V}^{ab}) = \hat{A}^{-1}\hat{\Sigma}(\hat{A}^{-1})'$.

Variance stabilizing transformations are examined here in terms of the quantity k which is defined by

$$k = -[\eta^a\eta^b\eta^c\,E\{\psi_a(X,\theta)\,\psi_b(X,\theta)\,\psi_c(X,\theta)\} + 4\eta^a\eta^b\delta^c\,E\{\psi_a(X,\theta)\,\psi_{b/c}(X,\theta)\} + 2\eta^a\delta^b\delta^c\,E\{\psi_{a/bc}(X,\theta)\} - 2\delta^a\delta^b g_{ab}]/\sigma^3, \tag{2.1}$$

where $\eta^a = A^{b/a}g_b$ and $\delta^a = -V^{ab}g_b$. In this expression and subsequently, we adopt the standard convention by which summation over the range $1, \ldots, q$ is assumed for any index that appears both as a subscript and as a superscript. Calculations of DiCiccio and Monti (2002) show that

$$k = \frac{n}{\sigma^3}E\{(\hat{\gamma} - \gamma)(\hat{\sigma}^2 - \sigma^2)\} \tag{2.2}$$

to error of order $O(n^{-1/2})$.

Our proposal for variance stabilization stems from equation (2.2) and the property that the quantity k is not parameterization invariant. Let $\phi = h(\gamma)$ be any monotonically increasing smooth transformation of γ. Then, in obvious notation,

$$k_\phi = k_\gamma + 2\sigma_\gamma\,h_{(2)}(\gamma)/h_{(1)}(\gamma), \tag{2.3}$$

where $h_{(j)}(\gamma) = \mathrm{d}^j h(\gamma)/\mathrm{d}\gamma^j$, $j = 1, 2$.

The quantity k can be estimated by its sample version

$$\hat{k} = \left[\frac{1}{n} \sum \widehat{\mathrm{IF}}_i^3 + 4 \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \frac{1}{n} \sum \widehat{\mathrm{IF}}_i \hat{\eta}' \psi(X_i, \hat{\theta} + \varepsilon\hat{\delta}) \Big|_{\varepsilon=0} \right.$$
$$\left. - 2 \frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2} \frac{1}{n} \sum \hat{\eta}' \psi(X_i, \hat{\theta} + \varepsilon\hat{\delta}) \Big|_{\varepsilon=0} + 2 \frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2} g(\hat{\theta} + \varepsilon\hat{\delta}) \Big|_{\varepsilon=0} \right] \Big/ \hat{\sigma}^3,$$

where $\hat{\eta} = (\hat{\eta}^1, \ldots, \hat{\eta}^q)'$, $\hat{\delta} = (\hat{\delta}^1, \ldots, \hat{\delta}^q)'$, $\hat{\eta}^a = \hat{A}^{b/a} \hat{g}_b$, $\hat{\delta}^a = -\hat{V}^{ab} \hat{g}_b$ and $\widehat{\mathrm{IF}}_i = -\hat{\eta}' \psi(X_i, \hat{\theta}) = -\hat{\eta}^a \psi_a(X_i, \hat{\theta})$ is the empirical influence function for γ at X_i. Note that $\hat{\sigma}^2$ and $\hat{k}$ can be computed easily by using numerical differentiation; analytical expressions for the partial derivatives of $g(\theta)$ and $\psi(X, \theta)$ are unnecessary. Both σ^2 and k are of order $O(1)$, and the differences $\hat{\sigma}^2 - \sigma^2$ and $\hat{k} - k$ are both of order $O_p(n^{-1/2})$.

Analogously to equation (2.3), the quantity $\hat{k}$ satisfies the transformation rule

$$\hat{k}_\phi = \hat{k}_\gamma + 2\hat{\sigma}_\gamma h_{(2)}(\hat{\gamma}) / h_{(1)}(\hat{\gamma}), \tag{2.4}$$

which can be used to find a data-dependent transformation $\phi = h(\gamma)$ satisfying $\hat{k}_\phi = 0$.

2.2. The transformation

Formula (2.2) permits a more formal understanding of the graphical approach to variance stabilization of Davison and Hinkley (1997), sections 3.9 and 5.2. The distribution of the bootstrap random variable $(\hat{\gamma}^*, \hat{\sigma}^{*2})$ has

$$\hat{k} = \frac{n}{\hat{\sigma}^3} E\{(\hat{\gamma}^* - \hat{\gamma})(\hat{\sigma}^{*2} - \hat{\sigma}^2)\}$$

to error of order $O(n^{-1/2})$ given the sample values. If $\hat{k} = 0$, then $\hat{\gamma}^*$ and $\hat{\sigma}^{*2}$ are nearly uncorrelated. A transformation for which the variance parameter plot shows no systematic linear trend would have $\hat{k}$ near 0.

Our proposal, therefore, is to find, from some suitable class of transformations, a reparameterization $\phi = h(\gamma)$ for which $\hat{k}_\phi = 0$. The transformation rule (2.4) would require that the reparameterization satisfies

$$\frac{\mathrm{d}[\log\{h_{(1)}(\gamma)\}]}{\mathrm{d}\gamma} \Big|_{\gamma=\hat{\gamma}} = -\frac{1}{2} \frac{\hat{k}_\gamma}{\hat{\sigma}_\gamma}. \tag{2.5}$$

The choice of parameterization is entirely empirical, depending only on the data. In subsequent sections, we show, however, that this proposal does indeed have variance stabilization properties in both parametric and nonparametric problems.

Convenient classes of transformations to consider for deriving variance stabilizing reparameterizations are provided by the Box–Cox transformations (Box and Cox, 1964). If the parameter γ is positive, then the class of Box–Cox transformations

$$h(\gamma) = \frac{\gamma^\lambda - 1}{\lambda} \tag{2.6}$$

can be considered; equation (2.5) shows that $\hat{k}_\phi = 0$ when

$$\lambda = 1 - \tfrac{1}{2} \hat{\gamma} \hat{k}_\gamma / \hat{\sigma}_\gamma. \tag{2.7}$$

If the parameter γ is not restricted to be positive, then the Box–Cox transformations can be applied to $\exp(\gamma)$. This approach leads to the parameterization $\phi = h(\gamma)$ with

$$h(\gamma) = \frac{\exp(\lambda\gamma) - 1}{\lambda}, \tag{2.8}$$

and expression (2.5) shows that $\hat{k}_\phi = 0$ when

$$\lambda = -\tfrac{1}{2}\hat{k}_\gamma / \hat{\sigma}_\gamma. \tag{2.9}$$

More generally, a class of monotonically increasing transformations can be obtained by applying the Box–Cox transformations to an initial parameterization $f(\gamma)$, where $f(\gamma)$ is a positive function. In this approach, the parameterization $\phi = h(\gamma)$ is given by

$$h(\gamma) = \frac{\pm\{f(\gamma)^\lambda - 1\}}{\lambda}, \tag{2.10}$$

the negative sign being used when $f(\gamma)$ is monotonically decreasing. Expression (2.5) shows that $\hat{k}_\phi = 0$ when

$$\lambda = -\frac{1}{F_1(\hat{\gamma})}\left\{\frac{1}{2}\frac{\hat{k}_\gamma}{\hat{\sigma}_\gamma} + F_2(\hat{\gamma})\right\}, \tag{2.11}$$

where $F_1(\gamma) = \mathrm{d}[\log\{f(\gamma)\}]/\mathrm{d}\gamma$ and $F_2(\gamma) = \mathrm{d}\{\log|F_1(\gamma)|\}/\mathrm{d}\gamma$.

2.3. Further remarks

If the goal is to identify a parameterization for which the variance parameter plot shows no systematic trend, and in particular no linear trend, then it might seem more natural to work with ρ, the correlation between $\hat{\gamma}$ and $\hat{\sigma}_\gamma^2$, than to use the quantity k. The transformation rule for the correlation is

$$\rho_\phi = \frac{k_\phi}{(k_\phi^2 - k_\gamma^2 + \upsilon_\gamma)^{1/2}},$$

where $\upsilon_\gamma = n\,E\{(\hat{\sigma}_\gamma^2/\sigma_\gamma^2 - 1)^2\}$, and the condition that a monotonically increasing transformation $\phi = h(\gamma)$ must satisfy for $\hat{\rho}_\phi = 0$ is identical to expression (2.5). For this reason, and since, as discussed in Section 1, the quantity $\hat{k}$ has a meaningful interpretation in terms of measuring the change in relative variance locally near $\hat{\gamma}$, it is convenient to work directly with $\hat{k}$ instead of using the correlation coefficient.

Although the development here is in terms of independent and identically distributed observations, only the independence assumption is crucial. In many situations, such as regression models, the ψ-function varies across observations, so that the M-estimator $\hat{\theta}$ satisfies $\Sigma\,\psi^i(X_i, \hat{\theta}) = 0$, where

$$\psi^i(X_i, \theta) = (\psi_1^i(X_i, \theta), \ldots, \psi_q^i(X_i, \theta))'$$

is the ψ-function for observation X_i. Here, $A = (A_{a/b})$ and $\Sigma = (\Sigma_{ab})$ are defined in obvious notation to have entries

$$A_{a/b} = E\{n^{-1}\textstyle\sum \psi_{a/b}^i(X_i, \theta)\}$$

and

$$\Sigma_{ab} = E\{n^{-1}\textstyle\sum \psi_a^i(X_i, \theta)\,\psi_b^i(X_i, \theta)\}.$$

The results extend in a direct manner to such situations by substituting $\psi^i(X_i, \theta)$ for $\psi(X_i, \theta)$ in the calculation of $\hat{\sigma}^2$ and $\hat{k}$.

The smooth function of means framework (Hall (1992), section 2.4), where $X=(X^1,\ldots,X^q)'$ and the parameter of interest $\gamma=g(\mu)$ is a smooth function of the vector mean $\mu=E(X)$, is a special case of M-estimation with $\psi(X,\mu)=X-\mu$. In this situation, the expressions for k and $\hat{k}$ simplify considerably. In particular, $V=\Sigma=E\{(X-\mu)(X-\mu)'\}$, $\sigma^2=g'_{(1)}\Sigma g_{(1)}$, $\eta=-g_{(1)}$ and $\delta=-\Sigma g_{(1)}$. Hence,

$$k=[E\{(X-\mu)'g_{(1)}\}^3+2\delta' g_{(2)}\delta]/\sigma^3,$$

where $g_{(2)}=\partial^2 g(\mu)/\partial\mu\,\partial\mu'$; furthermore,

$$\hat{k}=\left\{n^{-1}\sum\widehat{\mathrm{IF}}_i^3+2\frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2}g(\hat{\theta}+\varepsilon\hat{\delta})\bigg|_{\varepsilon=0}\right\}\Big/\hat{\sigma}^3,$$

where $\widehat{\mathrm{IF}}_i=(X_i-\bar{X})'\hat{g}_{(1)}$, $\hat{\delta}=-\hat{\Sigma}\hat{g}_{(1)}$, $\hat{\sigma}^2=\hat{g}'_{(1)}\hat{\Sigma}\hat{g}_{(1)}$, $\bar{X}=n^{-1}\Sigma X_i$, $\hat{g}_{(1)}=g_{(1)}(\hat{\theta})$ and $\hat{\Sigma}=n^{-1}\Sigma(X_i-\bar{X})(X_i-\bar{X})'$.

Although the transformation method that was proposed in Section 2.2 is not generally parameterization invariant, it is insensitive to some choices of initial parameterization. If the parameter γ is positive, then the Box–Cox class (2.6) with the exponent λ_γ given by equation (2.7) can be used; alternatively, the class (2.10) could be used with initial parameterization $f(\gamma)=\gamma^r$ for any power r, in which case the exponent that is obtained from equation (2.11) is $\lambda_{\gamma^r}=\lambda_\gamma/r$. The transformations that are obtained this way from classes (2.6) and (2.10) would differ only by a scalar multiple, so they would perform identically in terms of variance stabilization. Similarly, if the parameter γ takes negative values, then the class (2.8) is available or, alternatively, the class (2.10) can be used with initial parameterization $f(\gamma)=\exp(c\gamma)$ for any non-zero scalar c. For this choice of initial parameterization, the exponent that is given by equation (2.11) is λ_γ/c, where λ_γ is the exponent (2.9) for use with the class (2.8), so the transformations that are obtained from classes (2.8) and (2.10) would perform identically.

The lack of parameterization invariance requires that care should be taken when choosing a parameterization in which the method proposed is to be applied. In particular, the method seems to work better when used with a parameterization that has no range constraints beyond the positivity of γ that is required for the Box–Cox class (2.6) or, more generally, the positivity of the function $f(\gamma)$ that is required for class (2.10). If the parameter γ has an upper limit c and $\hat{\gamma}$ is close to c, then using the initial parameterization $f(\gamma)=c-\gamma$ is likely to produce better results than using γ directly; similarly, if $\hat{\gamma}$ is close to a lower limit c, then $f(\gamma)=\gamma-c$ is preferred to γ. Issues relating to the initial choice of parameterization are examined further in Section 3.4.

2.4. Examples

2.4.1. *Example 1: air-conditioning data failure ratio problem of Davison and Hinkley (1997)*

In the first example, there are two samples of sizes $n_1=12$ and $n_2=24$ drawn from two populations having positive means μ_1 and μ_2, and the parameter of interest is $\gamma=\mu_2/\mu_1$. Because inference is based on two independent samples, of different sizes, from two distributions, this problem is not in the smooth function framework, but it can be handled by the more general M-estimation approach of Section 2.3. Davison and Hinkley (1997), page 219, demonstrated the effectiveness of the logarithmic transformation for these data to stabilize variance. Formula (2.7) yields $\lambda_\gamma=-0.08949$, which agrees closely with their choice of transformation.

2.4.2. *Example 2: city population data problem of Davison and Hinkley (1997)*

In the second example, there is a sample of size $n=10$ from a bivariate distribution having mean (μ_1,μ_2), where μ_1 and μ_2 are both positive. The parameter of interest is $\gamma=\mu_2/\mu_1$. In contrast

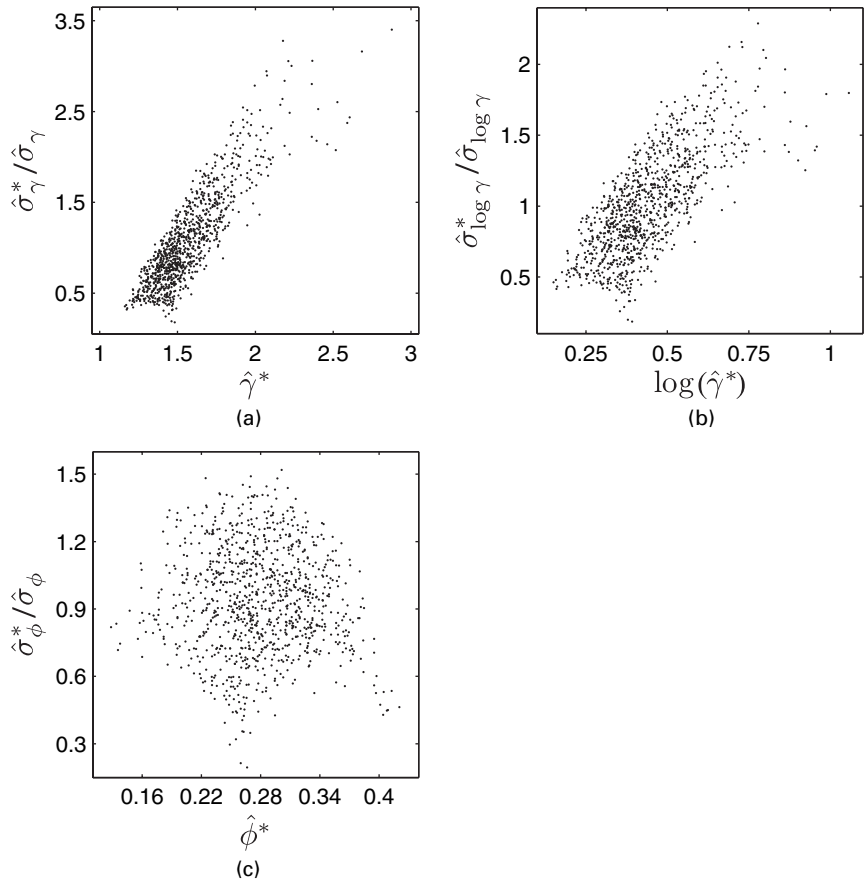

Fig. 1. Standard deviation *versus* parameter plots for ratio of means $\gamma = \mu_2/\mu_1$ ($B = 999$): (a) γ; (b) log(γ); (c) ϕ, the Box–Cox transformation applied to γ

with example 1, this problem is in the smooth function framework. Davison and Hinkley (1997), page 113, used this example to demonstrate the Tibshirani (1988) method of variance stabilization based on smoothing the variance parameter plot. Formula (2.7) yields $\lambda_\gamma = -2.1275$. Fig. 1 shows standard deviation *versus* parameter plots based on $B = 999$ bootstrap samples: Fig. 1(a) shows $\hat{\sigma}^*_\gamma/\hat{\sigma}_\gamma$ *versus* $\hat{\gamma}^*$; Fig. 1(b) shows $\hat{\sigma}^*_{\log(\gamma)}/\hat{\sigma}_{\log(\gamma)}$ *versus* $\log(\hat{\gamma}^*)$; Fig. 1(c) shows $\hat{\sigma}^*_\phi/\hat{\sigma}_\phi$ *versus* $\hat{\phi}^*$, where ϕ is the Box–Cox parameterization based on γ. The Box–Cox transformation is clearly effective for reducing trend in the plot; the sample correlations for these plots are 0.868, 0.738 and -0.051, respectively for Figs 1(a), 1(b) and 1(c). A comparison with Fig. 3.11 of Davison and Hinkley (1997) shows that the Box–Cox transformation works comparably with Tibshirani's method. Based on 5 million bootstrap samples, the 95% confidence interval for γ from the bootstrap t applied on the original scale is $[1.245, 2.096]$, applied on the log-scale is $[1.232, 2.147]$ and applied on the Box–Cox scale is $[1.204, 2.367]$. Based on 10 000 bootstrap samples, each with 500 subsamples used for variance estimation, Tibshirani's method yields the interval $[1.139, 2.470]$. The Box–Cox transformation appears to be making a similar adjustment to that of the Tibshirani method; the upper end point is noticeably larger, whereas the lower end point is smaller.

Davison and Hinkley (1997), section 3.10.2, discussed the use of bootstrap scatterplots to

determine a transformation $h(\gamma)$ such that $h(\hat{\gamma}^*)$ is close to its linear approximation. They considered the Box–Cox family of transformations (2.6) in connection with the city population data problem and showed that linearity is almost achieved when $\lambda = -2$, which is very close to the value $\lambda_\gamma = -2.1275$ that is obtained from equation (2.7). Davison and Hinkley (1997), example 3.25, argued that use of this transformation improves the normal approximation to the distribution of the Studentized statistic.

2.4.3. Example 3: cd4 count data analysed by DiCiccio and Efron (1996)

The data set for the third example consists of a sample of size $n = 20$ drawn from a bivariate distribution. If the parameter of interest γ is the largest eigenvalue of the covariance matrix of the underlying bivariate population, formula (2.7) yields $\lambda = 0.52032$. Canty *et al.* (1996) demonstrated that the square-root transformation is variance stabilizing in this situation, which agrees with the present finding.

Now suppose that the parameter of interest is ρ, the correlation of the bivariate population; here, $\hat{\rho} = 0.72317$ and $\hat{\sigma} = 0.35547$. If the Box–Cox transformation is applied to ρ directly, then formula (2.7) yields $\lambda_\rho = 3.0237$. If $\hat{\rho}$ is large and positive, then it is reasonable to apply the Box–Cox transformation in terms of the parameterization $1 - \rho$, which has the benefit of mapping the limit $\rho = 1$ to 0. Formula (2.7) yields $\lambda_{1-\rho} = 0.22531$. Similarly, it is reasonable to apply the Box–Cox transformation to

$$\xi = (1 + \rho)/(1 - \rho),$$

whose range is the entire positive axis, and equation (2.7) yields $\lambda_\xi = -0.05571$, indicating that Fisher's transformation is appropriate. Fig. 2 shows standard deviation *versus* parameter plots based on $B = 999$ bootstrap samples: Fig. 2(a) shows $\hat{\sigma}^*_\rho/\hat{\sigma}_\rho$ *versus* $\hat{\rho}^*$; Fig. 2(b) shows $\hat{\sigma}^*_{\phi_\rho}/\hat{\sigma}_{\phi_\rho}$ *versus* $\hat{\phi}^*_{\phi_\rho}$, where ϕ_ρ is the Box–Cox transformation applied to ρ; Fig. 2(c) shows $\hat{\sigma}^*_{\phi_{1-\rho}}/\hat{\sigma}_{\phi_{1-\rho}}$ *versus* $\hat{\phi}^*_{1-\rho}$, where $\phi_{1-\rho}$ is the Box–Cox transformation applied to $1 - \rho$; Fig. 2(d) shows $\hat{\sigma}^*_{\phi_\xi}/\hat{\sigma}_{\phi_\xi}$ *versus* $\hat{\phi}^*_\xi$, where ϕ_ξ is the Box–Cox transformation applied to ξ. The sample correlations for these plots are -0.798, 0.073, -0.062 and 0.064 respectively.

To illustrate the effect in this example of the choice of initial parameterization for the Box–Cox transformation, Table 1 reports nominal 90% and 99% equitailed confidence intervals for ρ derived from the standard normal approximation to the distributions of various Studentized pivots. The Studentized pivots are based on the original parameterization ρ and parameterizations that are obtained by applying the Box–Cox transformation to ρ, $\exp(\rho)$, $1 - \rho$, $e - \exp(\rho)$ and $(1 + \rho)/(1 - \rho)$. For the 90% confidence intervals, the Box–Cox parameterizations all yield similar intervals, and these intervals are noticeably different from the interval that is based on ρ. As the coverage level increases, however, differences emerge between the intervals that are obtained by applying the Box–Cox transformation to ρ and $\exp(\rho)$ on the one hand and those that are obtained by applying the transformation to $1 - \rho$, $e - \exp(\rho)$, and $(1 + \rho)/(1 - \rho)$ on the other.

3. Maximum likelihood estimation in parametric models

3.1. General background

Consider a parametric model that is indexed by $\theta = (\theta^1, \ldots, \theta^q)$, and let $l(\theta; X)$ denote the log-likelihood function that is based on a single observation X. Take $\psi_a(X, \theta) = l_a(\theta; X)$, where $l_a(\theta; X) = \partial l(\theta; X)/\partial\theta^a$ $(a = 1, \ldots, q)$, so that M-estimation corresponds to maximum likelihood estimation. It is shown in Appendix A that

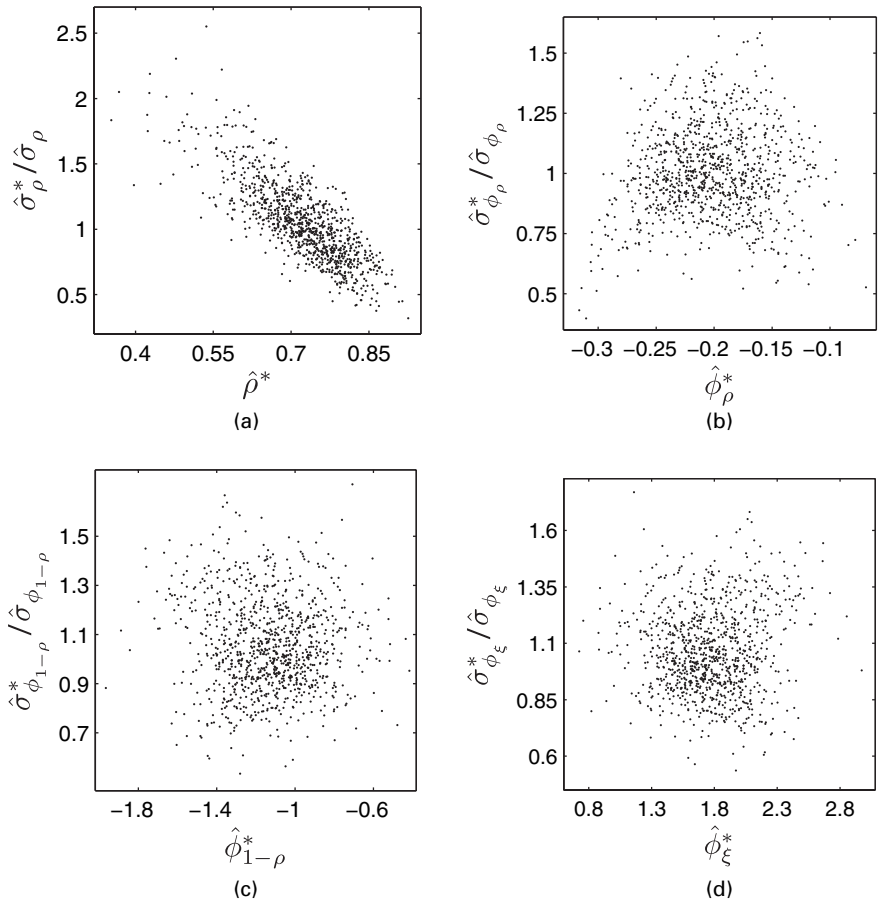

Fig. 2. Standard deviation *versus* parameter plots for correlation coefficient ρ ($B = 999$): (a) ρ; (b) ϕ_ρ, the Box–Cox transformation applied to ρ; (c) $\phi_{1-\rho}$, the Box–Cox transformation applied to $1 - \rho$; (d) ϕ_ξ, the Box–Cox transformation applied to $\xi = (1 + \rho)/(1 - \rho)$

Table 1. Equitailed confidence intervals for example 3, the correlation coefficient ρ, based on standard normal approximations to the distributions of Studentized pivots†

Parameterization	λ	*Confidence intervals for the following nominal two-sided coverages:*	
		90%	*99%*
ρ		[0.5924, 0.8539]	[0.5184, 0.9279]
Box–Cox, ρ	3.02369	[0.5567, 0.8354]	[0.3809, 0.8873]
Box–Cox, $\exp(\rho)$	2.79838	[0.5604, 0.8346]	[0.4191, 0.8850]
Box–Cox, $1 - \rho$	0.22531	[0.5664, 0.8320]	[0.4513, 0.8767]
Box–Cox, $e - \exp(\rho)$	0.42641	[0.5649, 0.8326]	[0.4440, 0.8787]
Box–Cox, $(1+\rho)/(1-\rho)$	−0.05571	[0.5678, 0.8314]	[0.4582, 0.8750]

†Point estimate $\hat{\rho} = 0.7232$.

$$k = -\{\eta^a \eta^b \eta^c (\partial \lambda_{bc}/\partial \theta^a) - 2\eta^a \eta^b g_{ab}\}/\sigma^3, \tag{3.1}$$

where $\lambda_{ab} = E(l_{ab})$, $l_{ab} = \partial^2 l(\theta; X)/\partial\theta^a\,\partial\theta^b$, $\eta^a = \lambda^{ab} g_b$ and $\sigma^2 = -\lambda^{ab} g_a g_b$, and (λ^{ab}) is the $q \times q$ matrix inverse of (λ_{ab}). In the parametric context, k can be approximated to error of order $O_p(n^{-1/2})$ by evaluating expression (3.1) at the maximum likelihood estimator $\hat{\theta}$, so take $\hat{k} = k(\hat{\theta})$. Similarly, let $\hat{\sigma}^2 = -\hat{\lambda}^{ab}\hat{g}_a\hat{g}_b$.

3.2. Scalar parameter models

Consider the case $q = 1$ so that $\theta = \theta^1$, and suppose that the parameter of interest is $\gamma = \theta$. In this situation, $\sigma^2 = -\lambda^{11} = -1/\lambda_{11}$; expression (3.1) yields

$$\begin{aligned} k &= \frac{\mathrm{d}\lambda_{11}}{\mathrm{d}\theta}(-\lambda^{11})^{3/2} \\ &= \frac{1}{\sigma}\frac{\mathrm{d}\sigma^2}{\mathrm{d}\theta}, \end{aligned} \tag{3.2}$$

so expression (1.1) is readily seen to hold. Approximate variance stabilization is achieved locally near $\hat{\theta}$ if $\hat{k} = 0$, since that condition implies $\mathrm{d}\sigma^2/\mathrm{d}\theta|_{\theta=\hat{\theta}} = 0$. However, if $k = 0$ for all θ, then global variance stabilization is achieved as, in that case, $\mathrm{d}\sigma^2/\mathrm{d}\theta = 0$ for all θ, i.e. σ^2 is constant.

3.2.1. *Example 4: binomial distribution having probability p*

The variance stabilizing transformation for this example model is $\sin^{-1}(p^{1/2})$. To apply the Box–Cox transformation in terms of p, note that $\sigma_p^2 = p(1-p)$; expression (3.2) yields

$$k_p = (1-2p)/\{p(1-p)\}^{1/2},$$

from which expression (2.7) produces $\lambda_p = (1-\hat{p})/2$. Alternatively, it is reasonable to apply the Box–Cox transformation in terms of the odds $\omega = p/(1-p)$, since the range of ω is the positive axis. Note that

$$\begin{gathered} \sigma_\omega^2 = \omega(1+\omega)^2 = p/(1-p)^3, \\ k_\omega = (1+3\omega)/\sqrt{\omega}, \\ \lambda_\omega = (1-\hat{\omega})/(2+2\hat{\omega}). \end{gathered}$$

Fig. 3 shows plots of the variance stabilizing transformation (full curve), $\phi_\omega = (\omega^\lambda - 1)/\lambda$ (broken curve) and $\phi_p = (p^\lambda - 1)/\lambda$ (chain curve). Fig. 3(a) is for $\hat{p} = 0.7$ and Fig. 3(b) is for $\hat{p} = 0.9$. The curves in Fig. 3 are standardized to take value 0 and to have slope 1 at $\hat{p}$. The Box–Cox transformation based on the odds ω is extremely effective, even for $\hat{p} = 0.9$.

3.2.2. *Example 5: bivariate normal distribution having known means and variances and unknown correlation coefficient ρ*

In example 5 the maximum likelihood estimator is not the usual sample correlation coefficient. Assume that the means are 0 and the variances are 1; from observations (X_i^1, X_i^2), $i = 1, \ldots, n$, the maximum likelihood estimator $\hat{\rho}$ is found as a root of a cubic equation, namely

$$S_{12} + (n - S_{11} - S_{22})\hat{\rho} + S_{12}\hat{\rho}^2 - n\hat{\rho}^3 = 0,$$

where $S_{rs} = \Sigma X_i^r X_i^s$. Furthermore, $\sigma_\rho^2 = (1-\rho^2)^2/(1+\rho^2)$, and the variance stabilizing transformation is

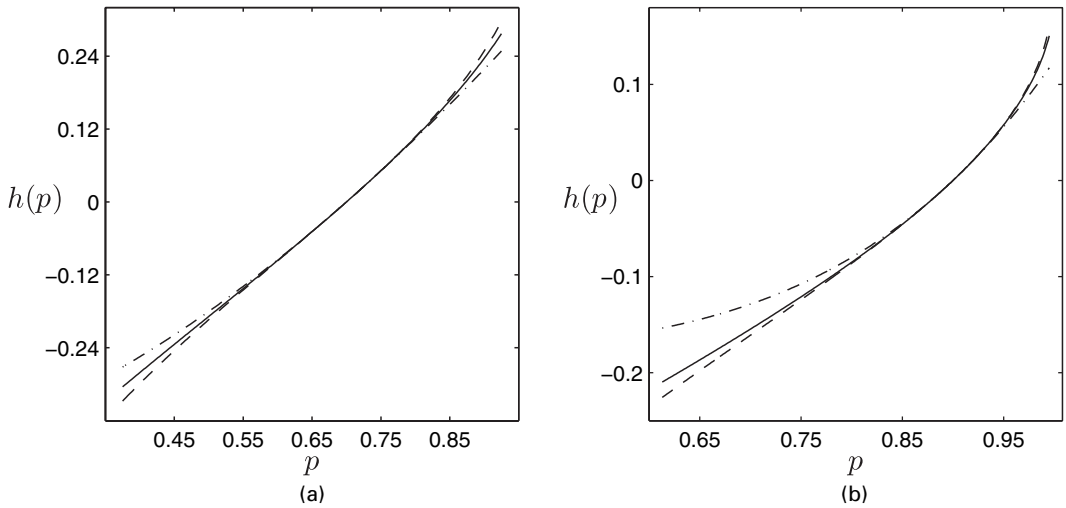

Fig. 3. Variance stabilizing transformations $h(p)$ for the binomial parameter p (———, exact variance stabilizing transformation; – – –, Box–Cox transformation applied to odds $\omega = p/(1-p)$; · – · – ·, Box–Cox transformation applied to p): (a) $\hat{p} = 0.7$; (b) $\hat{p} = 0.9$

$$\frac{1}{2}\log\left(\frac{1-\tau}{1+\tau}\right) - \frac{1}{2^{1/2}}\log\left(\frac{1-2^{1/2}\tau}{1+2^{1/2}\tau}\right), \qquad \tau = \frac{\rho}{(1+\rho^2)^{1/2}}.$$

If ρ is large and positive, then it is reasonable to apply the Box–Cox transformation in terms of the parameterization $1-\rho$, for which

$$k_{1-\rho} = 2\rho(3+\rho^2)/(1+\rho^2)^{3/2},$$
$$\lambda_{1-\rho} = (1-3\hat{\rho}+3\hat{\rho}^2-3\hat{\rho}^3)/(1-\hat{\rho}^4).$$

Note that $\lambda_{1-\rho} = 1$ when $\hat{\rho} = 0$. It is also sensible to apply the Box–Cox transformation in terms of Fisher's parameterization $\zeta = \tanh^{-1}(\rho)$, for which

$$k_\zeta = -2\rho(1-\rho^2)/(1+\rho^2)^{3/2},$$
$$\lambda_\zeta = \hat{\rho}(1-\hat{\rho}^2)/(1+\hat{\rho}^2),$$

where λ_ζ is obtained from equation (2.9). The identical transformation would be obtained by applying the Box–Cox method with λ given by equation (2.7) to the parameterization $(1+\rho)/(1-\rho)$. Fig. 4 shows plots of the exact variance stabilizing transformation (full curve), $\phi_\zeta = \{\exp(\lambda\zeta)-1\}/\lambda$ (broken curve) and $\phi_{1-\rho} = \{(1-\rho)^\lambda - 1\}/\lambda$ (chain curve): Fig. 4(a) is for $\hat{\rho} = 0.4$; Fig. 4(b) is for $\hat{\rho} = 0.7$; Fig. 4(c) is for $\hat{\rho} = 0.95$. The curves in Fig. 4 are standardized to take value 0 and to have slope 1 at $\hat{\rho}$. The Box–Cox transformation that is based on ζ provides excellent approximations to the true variance stabilizing transformation, and the Box–Cox transformation that is based on $1-\rho$ works extremely well for larger values of ρ.

3.3. Multiparameter models

Now consider the general case in which nuisance parameters are present. It follows from equation (3.1) that

$$k = -\frac{1}{\sigma^3}\eta^a \frac{\partial\sigma^2(\theta)}{\partial\theta^a}, \tag{3.3}$$

since $\sigma^2(\theta) = -\lambda^{bc} g_b g_c$ and $\partial\lambda^{bc}/\partial\theta^a = -\lambda^{bd}\lambda^{ce}\partial\lambda_{de}/\partial\theta^a$.

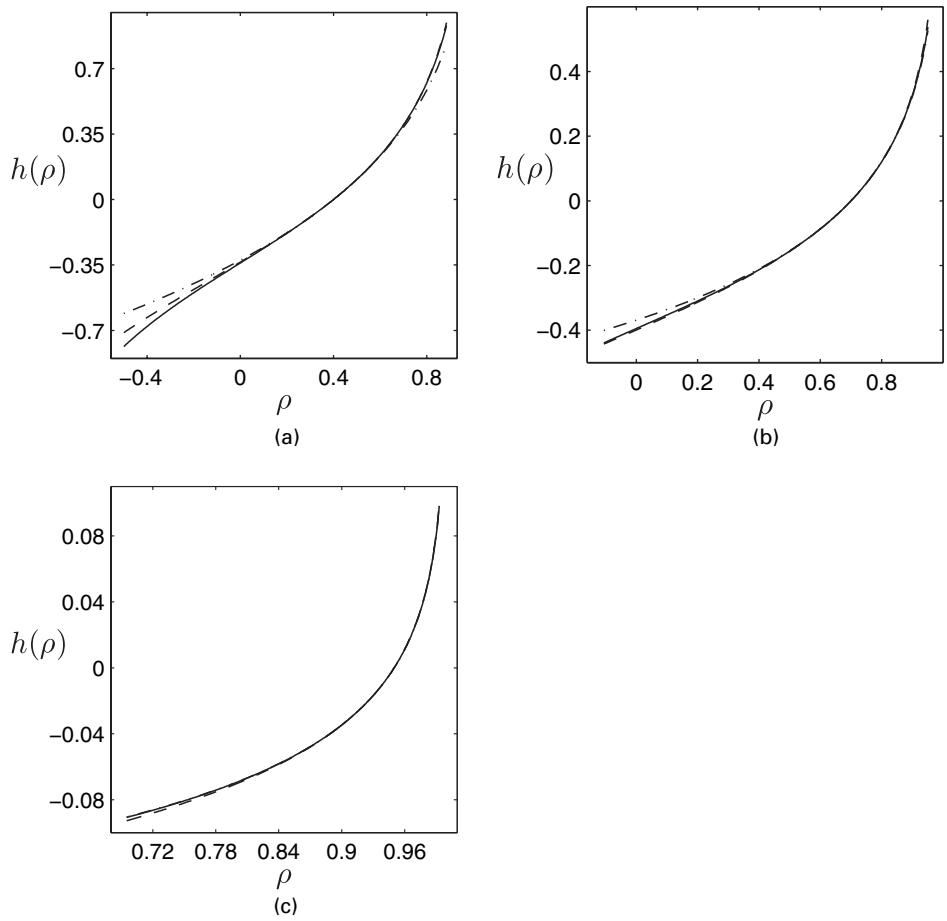

Fig. 4. Variance stabilizing transformations $h(\rho)$ for the correlation coefficient ρ (———, exact variance stabilizing transformation; – – –, Box–Cox transformation applied to Fisher's parameter $\zeta = \tanh^{-1}(\rho)$; · – · – ·, Box–Cox transformation applied to $1 - \rho$): (a) $\hat{\rho} = 0.4$; (b) $\hat{\rho} = 0.7$; (c) $\hat{\rho} = 0.95$

Let $\tilde{\theta} = \tilde{\theta}(\gamma)$ be the constrained maximum likelihood estimator of θ for a given value of γ, i.e. $\tilde{\theta}(\gamma)$ maximizes the log-likelihood function $\Sigma l(\theta; X_i)$ subject to the constraint $g(\theta) = \gamma$. Then $\tilde{\theta}(\gamma) = \{\tilde{\theta}^1(\gamma), \ldots, \tilde{\theta}^q(\gamma)\}$ is a curve through the parameter space for which $\tilde{\theta}(\hat{\gamma}) = \hat{\theta}$. Standard calculations show that

$$\left.\frac{\mathrm{d}\tilde{\theta}^a(\gamma)}{\mathrm{d}\gamma}\right|_{\gamma=\hat{\gamma}} = \frac{I^{ab}\hat{g}_b}{I^{bc}\hat{g}_b\hat{g}_c} = -\frac{1}{\hat{\sigma}^2}\hat{\eta}^a + O_p(n^{-1/2}), \qquad a = 1, \ldots, q, \tag{3.4}$$

where (I^{ab}) is the $q \times q$ inverse of (I_{ab}), the observed information matrix, whose elements are $I_{ab} = -\Sigma l_{ab}(\hat{\theta}; X_i)$; note that $I^{ab} = n^{-1}\hat{\lambda}^{ab} + O_p(n^{-3/2})$.

Now define $\tilde{\sigma}^2(\gamma)$ to be the asymptotic variance along the constrained maximum likelihood curve, so that $\tilde{\sigma}^2(\gamma) = \sigma^2\{\tilde{\theta}(\gamma)\}$. It follows from expression (3.4) that

$$\left.\frac{\mathrm{d}\tilde{\sigma}^2(\gamma)}{\gamma}\right|_{\gamma=\hat{\gamma}} = -\frac{1}{\hat{\sigma}^2}\hat{\eta}^a \left.\frac{\partial\sigma^2(\theta)}{\partial\theta^a}\right|_{\theta=\hat{\theta}} + O_p(n^{-1/2}), \tag{3.5}$$

and a comparison of expression (3.5) with expression (3.3) shows that expression (1.1) holds to error of order $O(n^{-1/2})$ given the sample values. For models in which the expected and observed information coincide, i.e. $I_{ab} = n\lambda_{ab}$ $(a, b = 1, \ldots, q)$, the error term on the right-hand side of equation (3.4) vanishes, and expression (1.1) holds exactly.

As the following example illustrates, the variance stabilizing transformations that are proposed here can be useful for improving the normal approximation to the distribution of the Studentized statistic in the parametric case.

3.3.1. Example 6: bivariate normal distribution having unknown means, variances and correlation coefficient ρ

In example 6, $\sigma_\rho^2 = (1-\rho^2)^2$, and Fisher's parameterization $\zeta = \tanh^{-1}(\rho)$ is variance stabilizing. As in example 5, if ρ is large and positive, it makes sense to apply the Box–Cox transformation to $1-\rho$, for which

$$k_{1-\rho} = 4\rho,$$
$$\lambda_{1-\rho} = (1-\hat{\rho})/(1+\hat{\rho}).$$

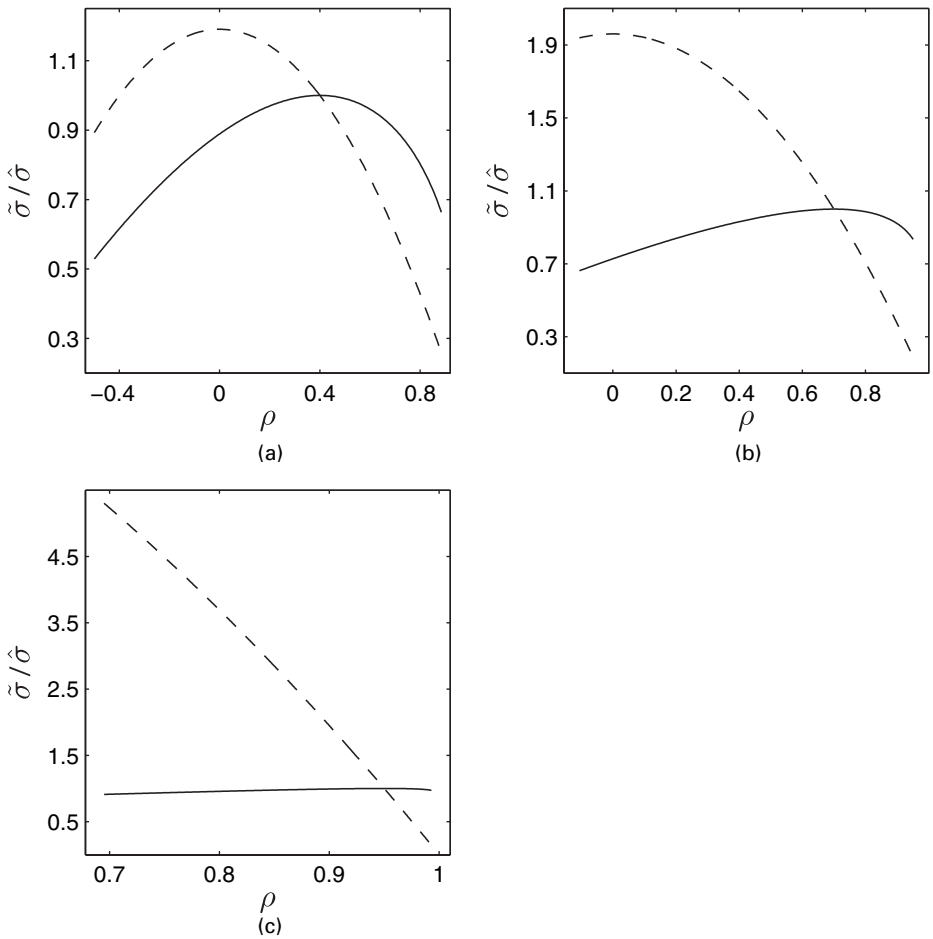

Fig. 5. Standard deviation plots for the correlation coefficient ρ (———, $\phi_{1-\rho}$, Box–Cox transformation applied to $1-\rho$; – – –, ρ): (a) $\hat{\rho} = 0.4$; (b) $\hat{\rho} = 0.7$; (c) $\hat{\rho} = 0.95$

Fig. 5 shows standard deviation *versus* parameter plots of $\tilde{\sigma}_\rho/\hat{\sigma}_\rho$ (broken curve) and $\tilde{\sigma}_{\phi_{1-\rho}}/\hat{\sigma}_{\phi_{1-\rho}}$ (full curve) *versus* ρ: Fig. 5(a) is for $\hat{\rho}=0.4$; Fig. 5(b) is for $\hat{\rho}=0.7$; Fig. 5(c) is for $\hat{\rho}=0.95$. The variance of $\hat{\rho}$ changes extremely rapidly, and the extent of variance stabilization that is afforded by the Box–Cox transformation $\phi_{1-\rho}=\{(1-\rho)^\lambda-1\}/\lambda$ is quite remarkable.

A simulation study was performed in this example to assess the efficacy of the method proposed. Table 2 shows properties of nominal 90% and 95% equitailed confidence intervals based on samples of size $n=20$ drawn from bivariate normal populations having $\rho=0.7$ and $\rho=0.9$. The confidence intervals are derived from the standard normal approximation to the distributions of Studentized statistics based on four parameterizations: the untransformed parameter ρ; the parameters that are obtained by using the Box–Cox transformation applied to $1-\rho$ with the exponents $\lambda_{1-\rho}=(1-\hat{\rho})/(1+\hat{\rho})$ and $\lambda^0_{1-\rho}=(1-\rho)/(1+\rho)$, where $\lambda^0_{1-\rho}$ is regarded as the 'true' value of $\lambda_{1-\rho}$; Fisher's parameter $\tanh^{-1}(\rho)$. The results that are reported in Table 2 are based on 100000 simulations.

The confidence intervals that are obtained from using ρ without transformation are unsatisfactory in terms of both their coverage errors and their expected lengths. The Box–Cox intervals are much better; their properties are closer to those of the Fisher intervals. It is interesting that the Box–Cox intervals that are based on $\lambda_{1-\rho}$, the 'estimated' exponent, are better than the intervals that are based on $\lambda^0_{1-\rho}$, the 'true' exponent. The end points of the Box–Cox confidence

Table 2. Simulated coverages and lengths of equitailed confidence intervals for the correlation coefficient ρ†

Parameterization	*Nominal two-sided coverage (%)*	*One-sided coverage error (%)*			*Two-sided coverage (%)*	*Interval length*	
		Lower limit	*Upper limit*	*Total*		*Mean*	*Standard deviation*
$\rho=0.7$							
Normal	90	−7.84	2.65	10.48	84.81	0.374	0.117
Box–Cox, $\lambda_{1-\rho}$		−2.13	0.05	2.18	87.92	0.385	0.115
Box–Cox, $\lambda^0_{1-\rho}$		−2.84	−0.84	3.68	86.32	0.386	0.119
Fisher		−0.66	0.91	1.57	90.25	0.411	0.119
Normal	95	−7.05	1.60	8.65	89.54	0.445	0.140
Box–Cox, $\lambda_{1-\rho}$		−1.48	0.01	1.49	93.53	0.465	0.136
Box–Cox, $\lambda^0_{1-\rho}$		−2.07	−0.98	3.05	91.95	0.466	0.144
Fisher		−0.46	0.42	0.89	94.96	0.493	0.139
$\rho=0.9$							
Normal	90	−9.22	4.14	13.37	84.92	0.144	0.063
Box–Cox, $\lambda_{1-\rho}$		−2.56	0.21	2.78	87.65	0.154	0.065
Box–Cox, $\lambda^0_{1-\rho}$		−2.76	−0.10	2.86	87.14	0.154	0.066
Fisher		−1.00	1.25	2.25	90.25	0.167	0.070
Normal	95	−8.26	2.37	10.63	89.11	0.172	0.075
Box–Cox, $\lambda_{1-\rho}$		−1.69	−0.01	1.70	93.30	0.189	0.079
Box–Cox, $\lambda^0_{1-\rho}$		−1.90	−0.36	2.26	92.74	0.189	0.081
Fisher		−0.53	0.68	1.21	95.15	0.205	0.084

†$\rho=0.7, 0.9$; $n=20$; simulation size, 100000. Normal intervals are derived from the Studentized version of ρ; $\lambda_{1-\rho}=(1-\hat{\rho})/(1+\hat{\rho})$; $\lambda^0_{1-\rho}=(1-\rho)/(1+\rho)$; the Fisher intervals are derived from the $\tanh^{-1}(\rho)$ transformation. If (θ_L, θ_U) is a nominal $1-\alpha$ confidence interval for θ, the upper and lower one-sided coverage errors are $P(\theta<\theta_U)-(1-\alpha/2)$ and $P(\theta>\theta_L)-(1-\alpha/2)$ respectively. The total one-sided coverage error is the sum of the absolute values of the upper and lower one-sided coverage errors.

intervals that are based on $\lambda_{1-\rho}$ are highly correlated to the end points of the Fisher intervals; the lower end points of the Box–Cox intervals tend to be larger than those of the Fisher intervals and the upper end points tend to be smaller. Thus, the Box–Cox intervals have lower coverage overall levels than the Fisher intervals and shorter expected lengths.

The values of $\lambda^0_{1-\rho}$ for $\rho = 0.7$ and $\rho = 0.9$ are 0.177 and 0.052 respectively. For $\rho = 0.7$, the mean and standard deviation of $\lambda_{1-\rho}$ are 0.191 and 0.098; when $\rho = 0.9$, the mean and standard deviation are 0.056 and 0.029. It appears that $\lambda_{1-\rho}$ estimates $\lambda^0_{1-\rho}$ satisfactorily in both instances; certainly, the variability in $\lambda_{1-\rho}$ is not detrimental to the performance of the confidence intervals.

3.4. *Choice of initial parameterization*

Suppose that $\phi = h(\gamma)$ is a reparameterization having $\hat{k}_\phi = 0$. Taking a further term in expansion (1.2) yields

$$\frac{\tilde{\sigma}^2_\phi(\phi)}{\hat{\sigma}^2_\phi} = 1 + \frac{1}{2} n^{-1} t^2 \frac{\mathrm{d}^2 \tilde{\sigma}^2_\phi(\phi)}{\mathrm{d}\phi^2}\bigg|_{\phi=\hat{\phi}} + O(n^{-3/2}), \tag{3.6}$$

which suggests that variance stabilization is more effectively achieved by the transformation $h(\gamma)$ when the second-order derivative on the right-hand side of expression (3.6) is small in magnitude. It can be shown that

$$\frac{\mathrm{d}^2 \tilde{\sigma}^2_\phi(\phi)}{\mathrm{d}\phi^2}\bigg|_{\phi=\hat{\phi}} = \frac{\mathrm{d}^2 \tilde{\sigma}^2_\gamma(\gamma)}{\mathrm{d}\gamma^2}\bigg|_{\gamma=\hat{\gamma}} - \hat{k}^2_\gamma + 2\hat{\sigma}^2_\gamma \frac{\mathrm{d}^2[\log\{h_{(1)}(\gamma)\}]}{\mathrm{d}\gamma^2}\bigg|_{\gamma=\hat{\gamma}}. \tag{3.7}$$

Expression (3.7) can elucidate the effect of the choice of initial parameterization $f(\gamma)$ for transformation (2.10) with λ given by equation (2.11). In this context,

$$\frac{\mathrm{d}^2[\log\{h_{(1)}(\gamma)\}]}{\mathrm{d}\gamma^2}\bigg|_{\gamma=\hat{\gamma}} = -\frac{1}{2}\frac{\hat{k}_\gamma}{\hat{\sigma}_\gamma} F_2(\hat{\gamma}) - F_2(\hat{\gamma})^2 + F_3(\hat{\gamma}), \tag{3.8}$$

where $F_3(\gamma) = \mathrm{d}^2\{\log|F_1(\gamma)|\}/\mathrm{d}\gamma^2 = \mathrm{d}F_2(\gamma)/\mathrm{d}\gamma$; thus, $\mathrm{d}^2\tilde{\sigma}^2_\phi(\phi)/\mathrm{d}\phi^2\big|_{\phi=\hat{\phi}}$ can be obtained by substituting expression (3.8) into expression (3.7). To simplify the notation in the following examples, denote $\mathrm{d}^2\tilde{\sigma}^2_\phi(\phi)/\mathrm{d}\phi^2\big|_{\phi=\hat{\phi}}$ by $\hat{K}_{f(\gamma)}$.

3.4.1. *Example 4 (continued)*

In the case of the binomial distribution having probability p,

$$\hat{K}_p = -(1-\hat{p})^{-1},$$
$$\hat{K}_{(1-p)} = -\hat{p}^{-1},$$
$$\hat{K}_{p/(1-p)} = 2.$$

Applying the method directly on the p-scale would work well for $\hat{p}$ near 0, but it would work poorly for $\hat{p}$ near 1. In Fig. 3, the chain curves, which correspond to ϕ_p, show noticeable curvature; the curvature is more pronounced for $\hat{p} = 0.9$ than it is for $\hat{p} = 0.7$, which reflects the behaviour of $\hat{K}_p$. Similarly, using the initial parameterization $f(p) = 1 - p$ would work well for $\hat{p}$ near 1, but it would work poorly for $\hat{p}$ near 0. In contrast, $\hat{K}_{\hat{p}/(1-\hat{p})}$ is not unbounded, so the initial parameterization $f(p) = p/(1-p)$ can be expected to produce good results regardless of the value of $\hat{p}$.

3.4.2. Example 6 (continued)
For the correlation coefficient problem,

$$\hat{K}_{\rho} = -8,$$
$$\hat{K}_{\exp(\rho)} = -4(1+\hat{\rho}^2),$$
$$\hat{K}_{1-\rho} = -4(1-\hat{\rho}),$$
$$\hat{K}_{1+\rho} = -4(1+\hat{\rho}).$$

Thus, using the initial parameterization $f(\rho) = \exp(\rho)$, i.e. using equation (2.8), would work well for $\hat{\rho}$ near 0, but its performance would deteriorate for $\hat{\rho}$ near 1 or -1. Using $f(\rho) = 1-\rho$ is advisable for $\hat{\rho}$ near 1, whereas $f(\rho) = 1+\rho$ is recommended for $\hat{\rho}$ near -1. In this case, $\log\{(1+\rho)/(1-\rho)\}$ is exactly variance stabilizing, so $\hat{K}_{(1+\rho)/(1-\rho)} = 0$.

The preceding two examples are typical, as the following general considerations show. Suppose that γ is a positive parameter having upper limit c, and suppose that $\tilde{\sigma}^2(\gamma)$ has the expansion

$$\tilde{\sigma}^2(\gamma) = (c-\gamma)d_1 + \frac{1}{2}(c-\gamma)^2 d_2 + \frac{1}{6}(c-\gamma)^3 d_3 + O\{(c-\gamma)^4\},$$

so that $\lim_{\gamma \to c}\{\tilde{\sigma}^2(\gamma)\} = 0$. It follows from expressions (3.6) and (3.7) that, to error of order $O\{(c-\hat{\gamma})^2)\}$,

$$\hat{K}_{\gamma} = \begin{cases} -\dfrac{1}{(c-\hat{\gamma})}d_1 - \dfrac{d_1}{c} - \dfrac{1}{2}d_2 - (c-\hat{\gamma})\left(\dfrac{d_1}{c^2} + \dfrac{d_2}{c} - \dfrac{1}{4}\dfrac{d_2^2}{d_1} - \dfrac{1}{6}d_3\right), & \text{if } d_1 \neq 0, \\ -d_2 - (c-\hat{\gamma})\left(\dfrac{d_2}{c} + \dfrac{1}{3}d_3\right), & \text{if } d_1 = 0, \end{cases}$$

whereas

$$\hat{K}_{c-\gamma} = \begin{cases} \dfrac{1}{2}d_2 + (c-\hat{\gamma})\left(\dfrac{1}{4}\dfrac{d_2^2}{d_1} + \dfrac{2}{3}d_3\right), & \text{if } d_1 \neq 0, \\ \dfrac{1}{6}(c-\hat{\gamma})d_3, & \text{if } d_1 = 0. \end{cases}$$

Example 4 has $d_1 = 1$ and example 6 has $d_1 = 0$. As $\hat{\gamma}$ approaches c, $\hat{K}_{\gamma}$ is unbounded when $d_1 \neq 0$ and tends to $-d_2$ when $d_1 = 0$; in contrast, $\hat{K}_{c-\gamma}$ tends to $d_2/2$ when $d_1 \neq 0$ and tends to 0 when $d_1 = 0$. Thus, in both cases, when $\hat{\gamma}$ is close to c, using $f(\gamma) = c-\gamma$ as initial parameterization in transformation (2.10) is recommended over using $f(\gamma) = \gamma$.

4. Nonparametric inference

The goal of the present section is to extend the previous results to the nonparametric framework; in particular, it is shown that the proposal for parameter transformation stabilizes variance appropriately along least favourable families (DiCiccio and Romano, 1990).

To specify one least favourable family, consider the family of probability distributions that is indexed by θ that arises in empirical likelihood (Owen, 1988). In this construction, we consider the probability distribution $p(\theta) = \{p^1(\theta), \ldots, p^n(\theta)\}$ that is defined on $X_1, \ldots, X_n$ where $p^i(\theta)$ is given by

$$p^i(\theta) = \frac{1}{n\{1+\nu^a(\theta)\,\psi_a(X_i,\theta)\}}, \qquad i = 1, \ldots, n,$$

and $\nu^1(\theta), \ldots, \nu^q(\theta)$ satisfy

$$\sum_{i=1}^{n} \psi_a(X_i, \theta)\, p^i(\theta) = 0, \qquad a = 1, \ldots, q. \tag{4.1}$$

The empirical likelihood function for θ is $L(\theta) = \Pi\, p^i(\theta)$, which is maximized at $\hat{\theta}$ satisfying $p^i(\hat{\theta}) = n^{-1}$ $(i = 1, \ldots, n)$. It follows from equation (4.1) that

$$\sum_{i=1}^{n} \psi(X_i, \hat{\theta}) = 0,$$

so $\hat{\theta}$ is the usual M-estimator.

A related family of probability distributions $p(\theta) = \{p^1(\theta), \ldots, p^n(\theta)\}$ that can be used for constructing a least favourable family is Efron's (1981) empirical exponential family; see also DiCiccio and Efron (1992). Empirical likelihood and empirical exponential family likelihood were discussed by Davison and Hinkley (1997), section 10.2.

Given a probability distribution $p(\theta)$, consider observations $X_1^*, \ldots, X_n^*$ that are drawn from $X_1, \ldots, X_n$ according to the probabilities $p^1(\theta), \ldots, p^n(\theta)$, and let $\hat{\theta}^*$ be the corresponding M-estimator of θ, i.e. $\Sigma\, \psi(X_i^*, \hat{\theta}^*) = 0$. Let $\hat{\gamma}^* = g(\hat{\theta}^*)$. Then, in somewhat abusive notation,

$$\mathrm{var}_{p(\theta)}\{n^{1/2}(\hat{\gamma}^* - \gamma)\} = \sigma^2(\theta) + O_p(n^{-1/2}),$$

where $g(\theta) = \gamma$, $\sigma^2(\theta) = V^{ab} g_a g_b$, $(V^{ab}) = A^{-1} \Sigma (A^{-1})'$, $A = (A_{a/b}(\theta))$ and $\Sigma = (\Sigma_{ab}(\theta))$, with

$$\begin{aligned} A_{a/b}(\theta) &= E_{p(\theta)}\{\psi_{a/b}(X^*, \theta)\} = \sum_{i=1}^{n} \psi_{a/b}(X_i, \theta)\, p^i(\theta), \\ \Sigma_{ab}(\theta) &= E_{p(\theta)}\{\psi_a(X^*, \theta)\, \psi_b(X^*, \theta)\} = \sum_{i=1}^{n} \psi_a(X_i, \theta)\, \psi_b(X_i, \theta)\, p^i(\theta). \end{aligned} \tag{4.2}$$

This notation is consistent with the previous notation in the sense that quantities $\hat{A}_{a/b}$, $\hat{\Sigma}_{ab}$, $\hat{V}^{ab}$ and $\hat{\sigma}^2$ that were defined previously coincide with $A_{a/b}(\hat{\theta})$, $\Sigma_{ab}(\hat{\theta})$, $V^{ab}(\hat{\theta})$ and $\sigma^2(\hat{\theta})$ respectively.

For a given value of γ, let $\tilde{\theta}(\gamma)$ be the constrained M-estimator of θ, i.e. $\tilde{\theta}(\gamma)$ is the point at which $L(\theta) = \Pi\, p^i(\theta)$ is maximized subject to the constraint $g(\theta) = \gamma$. The least favourable family is the probability distribution that is indexed by γ given by $p\{\tilde{\theta}(\gamma)\}$. Then $\tilde{\sigma}^2(\gamma) = \sigma^2\{\tilde{\theta}(\gamma)\}$ is the asymptotic variance along the least favourable family.

It is shown in Appendix A that expression (1.1) holds, i.e.

$$\hat{k} = \frac{1}{\hat{\sigma}} \frac{\mathrm{d}\tilde{\sigma}^2(\gamma)}{\mathrm{d}\gamma}\bigg|_{\gamma = \hat{\gamma}}. \tag{4.3}$$

Note that expression (4.3) offers an alternative method of computing $\hat{k}$ by numerical differentiation of $\tilde{\sigma}^2(\gamma)$ at $\hat{\gamma}$. However, this method would explicitly involve calculation of the least favourable family, whereas the method that is given in Section 2.1 which is based on numerical differentiation requires no constrained maximization. A referee has conjectured that, in principle, Tibshirani's (1988) method and the least favourable family approach of DiCiccio and Romano (1990) may be less sensitive to the choice of initial parameterization than is the proposal of this paper, and this topic deserves further investigation.

The maximizations that are necessary to compute the constrained M-estimation curve for least favourable families derived from the empirical likelihood and empirical exponential family

likelihood can be avoided by using Efron's (1987) exponential tilt family to approximate $\tilde{\theta}(\gamma)$ and $p\{\tilde{\theta}(\gamma)\}$. In the exponential tilt family that is indexed by a scalar β, the probability $p^i\{\tilde{\theta}(\gamma)\}$ is approximated by

$$\breve{p}^i(\beta) = \frac{\exp\{\beta(\widehat{\mathrm{IF}}_i)\}}{\sum \exp\{\beta(\widehat{\mathrm{IF}}_i)\}},$$

where $\widehat{\mathrm{IF}}_i = -\hat{\eta}^a \psi_a(X_i, \hat{\theta})$ is the empirical influence function for γ at X_i $(i = 1, \ldots, n)$, and $\tilde{\theta}(\gamma)$ is approximated by $\breve{\theta}(\beta)$ that satisfies

$$\frac{\sum \psi\{X_i, \breve{\theta}(\beta)\} \exp\{\beta(\widehat{\mathrm{IF}}_i)\}}{\sum \exp\{\beta(\widehat{\mathrm{IF}}_i)\}} = 0;$$

β is related to γ by the equation $\gamma(\beta) = g\{\breve{\theta}(\beta)\}$. The exponential tilt family was used in the following two examples.

4.1. *Example 2 (continued)*

Recall that, in example 2, the parameter of interest is $\gamma = \mu_2/\mu_1$. Fig. 6(a) shows standard deviation *versus* parameter plots: $\tilde{\sigma}_\gamma(\gamma)/\hat{\sigma}_\gamma$ *versus* γ (broken curve) and $\tilde{\sigma}_{\phi_\gamma}(\gamma)/\hat{\sigma}_{\phi_\gamma}$ *versus* γ (full curve) where ϕ_γ is the Box–Cox transformation based on γ. The extent of variance stabilization is striking.

4.2. *Example 3 (continued)*

Suppose in the cd4 count data that the parameter of interest is $\gamma = \sigma_1^2$, the variance of the first component. Fig. 6(b) shows standard deviation *versus* parameter plots: $\tilde{\sigma}_\gamma(\gamma)/\hat{\sigma}_\gamma$ *versus* γ (broken curve) and $\tilde{\sigma}_{\phi_\gamma}(\gamma)/\hat{\sigma}_{\phi_\gamma}$ *versus* γ (full curve) where ϕ_γ is the Box–Cox transformation based on γ. In this case, equation (2.7) yields $\lambda = 0.27371$.

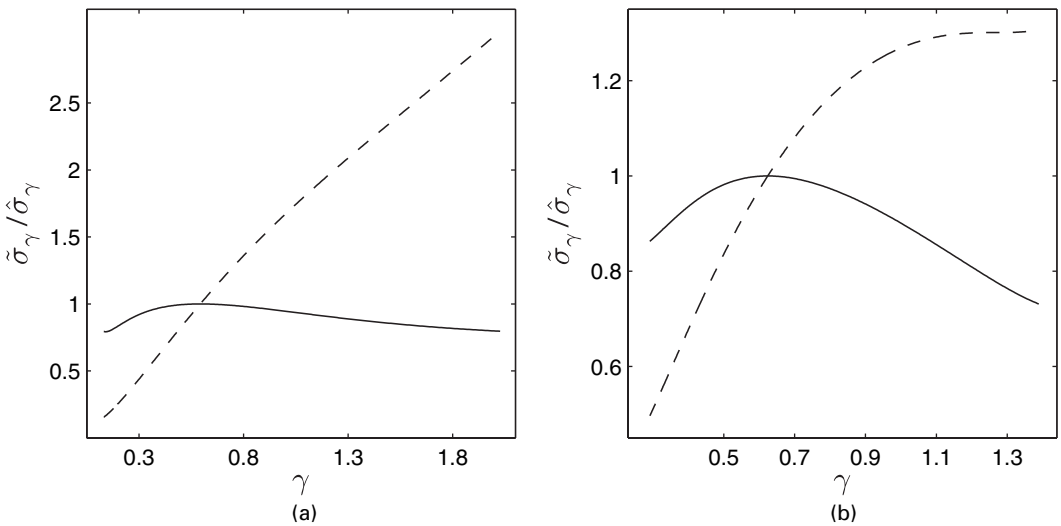

Fig. 6. Standard deviation plots (———, ϕ_γ, Box–Cox transformation applied to γ; – – –, γ): (a) parameter of interest $\gamma = \mu_1/\mu_2$, the ratio of means; (b) parameter of interest $\gamma = \sigma_1^2$, the variance of the first component

5. Discussion

In this paper, we have presented an automatic and general method to identify a variance stabilizing transformation when making inference about a scalar parameter. The procedure is applicable to parametric problems, though the work was motivated primarily by interest in nonparametric inference, where it is well known that commonly used inference methods such as the bootstrap work more effectively when the parameter of interest is essentially a location parameter, which can be induced by a variance stabilizing transformation. The current work was motivated specifically by interest in the properties of nonparametric profile likelihood, and in adjustments which reduce the bias of the profile score. Adjustments which are especially convenient for the nonparametric context are analogues of the parametric case adjustments that were devised by Cox and Reid (1993), which avoid the requirement of orthogonal nuisance parameters, but, unfortunately, lack the desirable property of parameterization invariance. Numerical evidence in the nonparametric case suggests that such adjustments will work best when applied in terms of a variance-stabilized parameterization.

Inference on the parameter γ of interest is typically based on the Studentized statistic $t = (\hat{\gamma} - \gamma)/\hat{\sigma}$, which is asymptotically normally distributed. In parametric contexts, the importance of skewness reduction in improving the accuracy of a normal approximation to the distribution of t is well appreciated: see, for example, Sprott (1980). In many parametric problems, variance stabilization does reduce skewness, but skewness reducing transformations are often more effective than variance stabilizing transformations for inducing normality. DiCiccio and Monti (2002) demonstrated how accurate inference about γ may be obtained when skewness is taken into account, by means of skewness reducing transformations. Their transformations were derived without explicit consideration of least favourable families. A natural issue is whether these transformations reduce skewness along the least favourable family, as we have demonstrated that the transformations in the present paper stabilize variance. In the current paper we have established a general machinery and framework of variance stabilization, and we now aim to investigate the effectiveness of variance stabilization in terms of inferential accuracy, compared with skewness reducing transformation, of procedures that are based on the normal approximation to the distribution of the Studentized statistic t, as well as procedures that are based on bootstrap methods and allied saddlepoint techniques. In many nonparametric cases, the two types of transformation are seen to be very close in examples, and it would be of interest to characterize cases where the two transformations might be expected to be close. More generally, it will be of interest to investigate the theoretical inferential properties, such as confidence interval coverage, of procedures which incorporate the empirical variance stabilization.

In parametric problems, interest often lies in conditional inference, given an ancillary statistic, rather than in unconditional inference. It will be of interest to examine the properties of the variance stabilizing transformations of the current paper from the perspective of conditional variance.

Acknowledgements

The authors are grateful to Nigar Hashimzade for providing the exact variance stabilizing transformation in example 5. The work of the first author was supported by a grant from the National Science Foundation; the work of the second author was supported by research funds from the University of Sannio. The authors thank the referees and Joint Editor for their helpful comments on the original version of the paper, and Anthony Davison for his hospitality at the Ecole Polytechnique Fédérale de Lausanne, during the preparation of the final version.

Appendix A

A.1. Proof of formula (3.1)

Let $l_{abc} = \partial^3 l(\theta; X)/\partial\theta^a\,\partial\theta^b\,\partial\theta^c$, and define $\lambda_{abc} = E(l_{abc})$, $\lambda_{a,b} = E(l_a l_b)$, $\lambda_{a,bc} = E(l_a l_{bc})$ and $\lambda_{a,b,c} = E(l_a l_b l_c)$ $(a, b, c = 1, \ldots, q)$. Note that the λs are of order $O(1)$. It follows from the identity $\lambda_{a,b} = -\lambda_{ab}$ that $A_{a/b} = -\Sigma_{ab} = \lambda_{ab}$, $\sigma^2 = -\lambda^{ab} g_a g_b$ and $\eta^a = \delta^a = \lambda^{ab} g_b$, where (λ^{ab}) is the $q \times q$ matrix inverse of (λ_{ab}). Hence, formula (2.1) reduces to

$$k = -\{\eta^a \eta^b \eta^c (\lambda_{a,b,c} + 4\lambda_{a,bc} + 2\lambda_{abc}) - 2\eta^a \eta^b g_{ab}\}/\sigma^3.$$

Furthermore, the identities $\lambda_{a,b,c} = -\lambda_{a,bc} - \lambda_{b,ac} - \lambda_{c,ab} - \lambda_{abc}$ and $\partial\lambda_{bc}/\partial\theta^a = \lambda_{a,bc} + \lambda_{abc}$ yield

$$\begin{aligned} k &= -\{\eta^a \eta^b \eta^c (\lambda_{a,bc} + \lambda_{abc}) - 2\eta^a \eta^b g_{ab}\}/\sigma^3 \\ &= -\{\eta^a \eta^b \eta^c (\partial\lambda_{bc}/\partial\theta^a) - 2\eta^a \eta^b g_{ab}\}/\sigma^3, \end{aligned}$$

as required.

A.2. Proof of formula (4.3)

By definition,

$$\left.\frac{\mathrm{d}\tilde{\sigma}^2(\gamma)}{\mathrm{d}\gamma}\right|_{\gamma=\hat{\gamma}} = \left.\frac{\partial\sigma^2(\theta)}{\partial\theta^a}\right|_{\theta=\hat{\theta}} \left.\frac{\mathrm{d}\tilde{\theta}^a(\gamma)}{\mathrm{d}\gamma}\right|_{\gamma=\hat{\gamma}}. \tag{A.1}$$

Calculations in the appendix of DiCiccio and Monti (2001) show that

$$\left.\frac{\mathrm{d}\tilde{\theta}^a(\gamma)}{\mathrm{d}\gamma}\right|_{\gamma=\hat{\gamma}} = \frac{\hat{V}^{ab}\hat{g}_b}{\hat{V}^{bc}\hat{g}_b\hat{g}_c} = -\frac{\hat{\delta}^a}{\hat{\sigma}^2}, \qquad a = 1, \ldots, q. \tag{A.2}$$

Moreover, differentiation of $\sigma^2(\theta) = V^{ab} g_a g_b$ yields

$$\begin{aligned} \frac{\partial\sigma^2(\theta)}{\partial\theta^a} &= -V^{bd} V^{ce} \frac{\partial V_{de}(\theta)}{\partial\theta^a} g_b g_c - 2g_{ab}\delta^b \\ &= -\frac{\partial V_{bc}(\theta)}{\partial\theta^a}\delta^b\delta^c - 2g_{ab}\delta^b, \end{aligned} \tag{A.3}$$

where $V^{-1} = (V_{ab}) = A'\Sigma^{-1}A$. Since $V_{ab} = \Sigma^{cd} A_{c/a} A_{d/b}$, it follows that

$$\frac{\partial V_{bc}(\theta)}{\partial\theta^a} = -\Sigma^{df}\Sigma^{eg}\frac{\partial\Sigma_{fg}(\theta)}{\partial\theta^a} A_{d/b}A_{e/c} + \Sigma^{de}\frac{\partial A_{d/b}(\theta)}{\partial\theta^a} A_{e/c} + \Sigma^{de} A_{d/b}\frac{\partial A_{e/c}(\theta)}{\partial\theta^a}. \tag{A.4}$$

Recall that $\delta^a = -V^{ab}g_b = -A^{a/c}\Sigma_{cd}A^{b/d}g_b$ and $\eta^a = A^{b/a}g_b$, so $\Sigma^{ab}A_{b/c}\delta^c = -\eta^a$. Thus, substitution of equation (A.4) into equation (A.3) and subsequent substitution of expressions (A.2) and (A.3) into expression (A.1) yield

$$\left.\frac{\mathrm{d}\tilde{\sigma}^2(\gamma)}{\mathrm{d}\gamma}\right|_{\gamma=\hat{\gamma}} = -\frac{1}{\hat{\sigma}^2}\left\{\left.\frac{\partial\Sigma_{bc}(\theta)}{\partial\theta^a}\right|_{\theta=\hat{\theta}}\hat{\delta}^a\hat{\eta}^b\hat{\eta}^c + 2\left.\frac{\partial A_{b/c}(\theta)}{\partial\theta^a}\right|_{\theta=\hat{\theta}}\hat{\delta}^a\hat{\eta}^b\hat{\delta}^c - 2\hat{g}_{ab}\hat{\delta}^a\hat{\delta}^b\right\}. \tag{A.5}$$

A key feature of any family of probability distributions $p(\theta) = \{p^1(\theta), \ldots, p^n(\theta)\}$ that yields a least favourable family $p\{\tilde{\theta}(\gamma)\}$ by evaluation along the constrained M-estimation curve is that

$$\left.\frac{\partial p^i(\theta)}{\partial\theta^a}\right|_{\theta=\hat{\theta}}\hat{\delta}^a = \frac{1}{n}\psi_b(X_i, \hat{\theta})\hat{\eta}^b, \qquad i = 1, \ldots, n.$$

Thus, differentiating the equations in expression (4.2) and substituting the results into expression (A.5) give

$$\left.\frac{\mathrm{d}\tilde{\sigma}^2(\gamma)}{\mathrm{d}\gamma}\right|_{\gamma=\hat{\gamma}} = -\frac{1}{\hat{\sigma}^2}\left\{\hat{\eta}^a\hat{\eta}^b\hat{\eta}^c\frac{1}{n}\sum_{i=1}^{n}\psi_a(X_i,\hat{\theta})\,\psi_b(X_i,\hat{\theta})\,\psi_c(X_i,\hat{\theta}) + 4\hat{\eta}^a\hat{\eta}^b\hat{\delta}^c\frac{1}{n}\sum_{i=1}^{n}\psi_a(X_i,\hat{\theta})\,\psi_{b/c}(X_i,\hat{\theta})\right.$$
$$\left. + 2\hat{\eta}^a\hat{\delta}^b\hat{\delta}^c\frac{1}{n}\sum_{i=1}^{n}\psi_{a/bc}(X_i,\hat{\theta}) - 2\hat{g}_{ab}\hat{\delta}^a\hat{\delta}^b\right\}$$
$$= \hat{\sigma}\hat{k},$$

which is the desired result.

References

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *J. R. Statist. Soc.* B, **26**, 211–243.

Canty, A. J., Davison, A. C. and Hinkley, D. V. (1996) Comment on 'Bootstrap confidence intervals' (by T. J. DiCiccio and B. Efron). *Statist. Sci.*, **11**, 214–219.

Cox, D. R. and Reid, N. (1993) A note on the calculation of adjusted profile likelihood. *J. R. Statist. Soc.* B, **55**, 467–471.

Daniels, H. E. and Young, G. A. (1991) Saddlepoint approximation for the studentized mean, with an application to the bootstrap. *Biometrika*, **78**, 169–179.

Davison, A. C. and Hinkley, D. V. (1988) Saddlepoint approximations in resampling methods. *Biometrika*, **75**, 417–431.

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

DiCiccio, T. J. and Efron, B. (1992) More accurate confidence intervals in exponential families. *Biometrika*, **79**, 231–245.

DiCiccio, T. J. and Efron, B. (1996) Bootstrap confidence intervals (with discussion). *Statist. Sci.*, **11**, 189–228.

DiCiccio, T. J., Martin, M. A. and Young, G. A. (1994) Analytical approximations to bootstrap distribution functions using saddlepoint methods. *Statist. Sin.*, **4**, 281–295.

DiCiccio, T. J. and Monti, A. C. (2001) Approximations to the profile empirical likelihood function for a scalar parameter in the context of *M*-estimation. *Biometrika*, **88**, 337–351.

DiCiccio, T. J. and Monti, A. C. (2002) Accurate confidence limits for scalar functions of vector *M*-estimands. *Biometrika*, **89**, 437–451.

DiCiccio, T. J. and Romano, J. P. (1990) Nonparametric confidence limits by resampling methods and least favorable families. *Int. Statist. Rev.*, **58**, 59–76.

DiCiccio, T. J. and Tibshirani, R. (1987) Bootstrap confidence intervals and bootstrap approximations. *J. Am. Statist. Ass.*, **82**, 163–170.

Efron, B. (1981) Nonparametric standard errors and confidence intervals (with discussion). *Can. J. Statist.*, **9**, 139–172.

Efron, B. (1987) Better bootstrap confidence intervals (with discussion). *J. Am. Statist. Ass.*, **82**, 171–200.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. London: Chapman and Hall.

Hall, P. (1992) *The Bootstrap and Edgeworth Expansion*. New York: Springer.

Hall, P. and Presnell, B. (1999) Intentionally biased bootstrap methods. *J. R. Statist. Soc.* B, **61**, 143–158.

Huber, P. J. (1981) *Robust Statistics*. New York: Wiley.

Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Sprott, D. A. (1980) Maximum likelihood in small samples: estimation in the presence of nuisance parameters. *Biometrika*, **67**, 515–523.

Tibshirani, R. J. (1988) Variance stabilization and the bootstrap. *Biometrika*, **75**, 433–444.