TESTS OF SEPARATE FAMILIES OF HYPOTHESES

D. R. COX
BIRKBECK COLLEGE, UNIVERSITY OF LONDON

1. Introduction

The aims of this paper are

(i) to stress the existence of a class of problems that have not received much attention in the literature;

(ii) to outline a general method, based on the Neyman-Pearson likelihood ratio, for tackling these problems;

(iii) to apply the general results to a few special cases.

Discussion of regularity conditions will not be attempted.

2. Some problems

The following are examples of the type of problem to be investigated. Throughout, Greek letters denote unknown parameters. It is assumed in the theoretical analysis that observed values of random variables are to be used to test one of the hypotheses, say H_f, and that high sensitivity is required for alternative hypotheses, H_g.

EXAMPLE 1. Let $Y_1, \cdots, Y_n$ be independent identically distributed random variables. Let $H_f^{(1)}$ be the hypothesis that their distribution function is log-normal with unknown parameter values, and let $H_g^{(1)}$ be the hypothesis that their distribution function is exponential with unknown parameter value. For remarks on the difficulty of distinguishing these distributions, see [11]. A. D. Roy [20] has given a likelihood ratio test for the similar example of discriminating between a normal and a log-normal distribution. Other related examples include that of testing a Weibull-type distribution against a gamma distribution, and of testing alternative forms of quantal dose-response curves.

EXAMPLE 2. Let $Y_1, \cdots, Y_n$ be independently normally distributed with constant variance and let $x_1, \cdots, x_n$ be given positive constants. Let $H_f^{(2)}$ be the hypothesis that

$$E(Y_i) = \alpha_1 + \alpha_2 x_i, \qquad i = 1, \cdots, n, \tag{1}$$

and let $H_g^{(2)}$ be the hypothesis that

$$E(Y_i) = \beta_1 + \beta_2 \log x_i, \qquad i = 1, \cdots, n. \tag{2}$$

EXAMPLE 3. Let $\mathbf{Y}$ denote a vector of n independently normally distributed random variables with constant variance and let $\mathbf{a}$, $\mathbf{b}$ be given matrices and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ vectors of unknown parameters, not necessarily with the same number of components. Let $H_f^{(3)}$ be the hypothesis that

$$E(\mathbf{Y}) = \mathbf{a}\boldsymbol{\alpha}, \tag{3}$$

and let $H_g^{(3)}$ be the hypothesis that

$$E(\mathbf{Y}) = \mathbf{b}\boldsymbol{\beta}. \tag{4}$$

Examples 2 and 3 are related to the problem, first considered by Hotelling [10], of choosing between independent variables in a regression relation; see [9], [24], and especially [25], for further discussion of Hotelling's results. As an instance of example 3, see [6] where, for data from a Latin square, a model is given quite different from the usual one.

EXAMPLE 4. Let $\mathbf{a}$ be a given matrix. Let $H_f^{(4)}$ be the hypothesis that $\log Y_1, \cdots, \log Y_n$ are independently normally distributed with constant variance and with

$$E(\log \mathbf{Y}) = \mathbf{a}\boldsymbol{\alpha} \tag{5}$$

and let $H_g^{(4)}$ be the hypothesis that $Y_1, \cdots, Y_n$ are independently normally distributed with constant variance and with

$$E(\mathbf{Y}) = \mathbf{a}\boldsymbol{\beta}. \tag{6}$$

Whereas examples 2 and 3 concern discrimination between alternative forms for the independent variable in a least squares problem, example 4 concerns alternative forms for the dependent variable. For instance, in a simple factorial experiment $H_f^{(4)}$ might be the hypothesis that $\log Y_i$ has zero two- and higher-factor interactions, $H_g^{(4)}$ the same hypothesis about Y_i. Rough methods for tackling these problems are almost certainly widely known, but the problems do not seem to have been considered theoretically.

Except in certain degenerate cases, as for example when $\alpha_1 = \beta_1 = 0$ in example 2, the hypotheses H_f and H_g are separate in the sense that an arbitrary simple hypothesis in H_f cannot be obtained as a limit of simple hypotheses in H_g. This contrasts with the usual situation in hypothesis testing.

There is a further group of problems that arise when one of the hypotheses, say H_g, involves a discontinuity in the model occurring at an unknown point in the sequence of observations. A sensible test criterion can be found by the methods of this paper, but the distributional problems seem difficult and will not be tackled in the present paper. The problems will, however, be recorded for reference.

EXAMPLE 5. Let $Y_1, \cdots, Y_n$ be independently normally distributed with constant variance and let $H_f^{(5)}$ be the hypothesis that

$$E(Y_i) = \alpha, \qquad i = 1, \cdots, n, \tag{7}$$

and let $H_g^{(5)}$ be the hypothesis that

$$E(Y_i) = \begin{cases} \beta_1, & i = 1, \cdots, \gamma, \\ \beta_2, & i = \gamma + 1, \cdots, n, \end{cases} \tag{8}$$

where γ is unknown.

This is different from the previous examples not only in involving a discontinuity, but also in that any simple hypothesis in $H_f^{(5)}$ can be obtained as a special case of $H_g^{(5)}$. The next examples illustrate a "continuous" hypothesis H_f and a separate "discontinuous" hypothesis H_g.

EXAMPLE 6. Let the assumptions of example 5 hold and let $H_g^{(6)}$ be the same as $H_g^{(5)}$, and let $H_f^{(6)}$ be the hypothesis that

$$E(Y_i) = \alpha_1 + i\alpha_2, \qquad i = 1, \cdots, n. \tag{9}$$

EXAMPLE 7. Let the Y_i and x_i be as in example 2. Let $H_f^{(7)}$ be the hypothesis that

$$E(Y_i) = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2, \qquad i = 1, \cdots, n, \tag{10}$$

whereas $H_g^{(7)}$ is the hypothesis that

$$E(Y_i) = \begin{cases} \beta_1 + \beta_2(x_i - \gamma), & x_i < \gamma, \\ \beta_1 + \beta_3(x_i - \gamma), & x_i \geqq \gamma. \end{cases} \tag{11}$$

The last example is relevant for testing whether data which obviously arise from a curved regression relation are better fitted by two straight lines at an angle than by a parabola. If $H_g^{(7)}$ is preferred to $H_f^{(7)}$, there is evidence for a discontinuous, or at any rate rapid, change in the regression relation. For instance x may be temperature, Y some physical or chemical property, for example, electrical resistance. It may be claimed on the basis of an apparent discontinuity in slope that a change in chemical structure occurs at temperature γ, and we may be concerned to examine the evidence for this claim.

Example 5 is a simplified version of the Lindisfarne scribes problem, in the general form of which the number of "change-points" is unknown; see [19], [21], and the contribution by Champernowne to the discussion of [19]. Page [16], [17] has considered a problem equivalent to example 5 with α, β_1, β_2 known; see also [3], [5], and for a relevant probabilistic result [8]. Quandt [17] has considered the maximum likelihood fitting of two straight lines at an angle.

3. A general formulation

Denote the vector random variable to be observed by $\mathbf{Y}$ and let H_f and H_g be respectively the hypotheses that the p.d.f. of Y is $f(\mathbf{y}, \boldsymbol{\alpha})$ and $g(\mathbf{y}, \boldsymbol{\beta})$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of unknown parameters, with $\boldsymbol{\alpha} \in \Omega_\alpha$ and $\boldsymbol{\beta} \in \Omega_\beta$. It is assumed, unless explicitly stated otherwise, that

(i) the families f and g are separate in the sense defined above;

(ii) the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ may be treated as varying continuously even when a component of say $\boldsymbol{\beta}$ is the serial number of the observation at which a discontinuity occurs;

(iii) the values of $\boldsymbol{\alpha}$, or $\boldsymbol{\beta}$, are interior to Ω_α, or Ω_β, so that the type of distribution problem discussed by Chernoff [4] is excluded.

There are several methods that might be used for these problems. First suppose that the problem is changed into one of discrimination, either H_f or H_g being true, and let there be prior probabilities $\tilde{\omega}_f$, $\tilde{\omega}_g$, with $\tilde{\omega}_f + \tilde{\omega}_g = 1$, for H_f, H_g, and conditionally on H_f, let $p_f(\boldsymbol{\alpha})$ be the prior p.d.f. of $\boldsymbol{\alpha}$, with $p_g(\boldsymbol{\beta})$ referring similarly to H_g. Then if the observed value of $\mathbf{Y}$ is denoted by $\mathbf{y}$, the posterior odds for H_f versus H_g are, by Bayes's theorem,

$$\frac{\tilde{\omega}_f \int_{\Omega_\alpha} f(\mathbf{y}, \boldsymbol{\alpha}) p_f(\boldsymbol{\alpha})\, d\boldsymbol{\alpha}}{\tilde{\omega}_g \int_{\Omega_\beta} g(\mathbf{y}, \boldsymbol{\beta}) p_g(\boldsymbol{\beta})\, d\boldsymbol{\beta}}. \tag{12}$$

In a decision problem account is taken of $w_f(\boldsymbol{\alpha})$ and $w_g(\boldsymbol{\beta})$, the losses due to wrongly rejecting H_f and H_g when $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are the true parameter points. This leads in the usual way to a decision rule of the form

$$\tilde{\omega}_f \int_{\Omega_\alpha} f(\mathbf{y}, \boldsymbol{\alpha}) p_f(\boldsymbol{\alpha}) w_f(\boldsymbol{\alpha})\, d\boldsymbol{\alpha} \gtrless \tilde{\omega}_g \int_{\Omega_\beta} g(\mathbf{y}, \boldsymbol{\beta}) p_f(\boldsymbol{\beta}) w_g(\boldsymbol{\beta})\, d\boldsymbol{\beta}. \tag{13}$$

Lindley [15] has obtained large-sample approximations to (12) and (13) by expansion about the maximum likelihood points $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$. His result leads to the following approximation to (12):

$$\frac{f(\mathbf{y}, \hat{\boldsymbol{\alpha}})}{g(\mathbf{y}, \hat{\boldsymbol{\beta}})} \frac{\tilde{\omega}_f}{\tilde{\omega}_g} \frac{(2\pi)^{d_f/2} p_f(\hat{\boldsymbol{\alpha}})\, \Delta_\alpha^{-1/2}}{(2\pi)^{d_g/2} p_g(\hat{\boldsymbol{\beta}})\, \Delta_\beta^{-1/2}}, \tag{14}$$

where d_f and d_g are the numbers of dimensions in the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and where Δ_α and Δ_β are the information determinants for estimating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from $\mathbf{Y}$.

If the prior probability distributions in (12) are known numerically, or if the approximation (14) can be used and the first factor is known numerically, this approach gives a general solution to the problem of discriminating between H_f and H_g, when no other distributions can arise. In the present paper we shal assume that numerical knowledge of prior distributions is not available.

A natural thing then is to introduce a statistic

$$r_{fg} \equiv e^{l_{fg}} = \begin{cases} \dfrac{\sup_{\boldsymbol{\alpha}\in\Omega_\alpha} f(\mathbf{y}, \boldsymbol{\alpha})}{\sup_{\boldsymbol{\beta}\in\Omega_\beta} g(\mathbf{y}, \boldsymbol{\beta})}, \\ f(\mathbf{y}, \hat{\boldsymbol{\alpha}})/g(\mathbf{y}, \hat{\boldsymbol{\beta}}), \end{cases} \tag{15}$$

a form of the Neyman-Pearson likelihood ratio. When l_{fg} is considered as a random variable, it is denoted by L_{fg}. In the common applications of the likelihood ratio $\Omega_\alpha \subset \Omega_\beta$, so that $L_{fg} < 0$; of course this inequality does not hold here.

The quantity l_{fg} is invariant under transformations of the observations and of the parameters.

There are several alternatives to (15) that might be considered. One is to replace $p_f(\boldsymbol{\alpha})$ and $p_g(\boldsymbol{\beta})$ in (12) by simple functions, for example by the invariant prior distributions of Jeffreys [12]. In the common cases where one or both parameters have infinite range, and the invariant prior distributions are therefore improper, the procedure is invalid because the normalizing constants for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not comparable. A second procedure, the OAAAA method (that is, obviously arbitrary and always admissible) of Barnard [3], is to take a small number of base points in Ω_α, each with the same formal prior probability under H_f, and similarly for H_g. This leads to a ratio of mean likelihoods rather than to a ratio of maximum likelihoods. If $\mathbf{Y}$ is a vector with a large number of independent components, this approach is equivalent to the use of (15), provided that for H_f and for H_g at least one of the base points is near enough to the maximum likelihood point. There are, in fact, many possible statistics that are asymptotically equivalent to l_{fg}; see [1], [22] for two such statistics when $\Omega_\alpha \subset \Omega_\beta$.

4. An alternative formulation

The conventional method of dealing with example 2, and with the more general example 3, is to set up a comprehensive model containing H_f and H_g as special cases. Thus for example 2, consider

$$E(Y_i) = \lambda_0 + \lambda_1 x_i + \lambda_2 \log x_i, \qquad i = 1, \cdots, n. \tag{16}$$

The λ can be estimated by least squares and if, say, λ_1 is very highly significant whereas λ_2 is not, a clear conclusion can be reached that the data agree better with H_f than with H_g.

An advantage of (16) is that it may lead to an adequate representation of the data when both H_f and H_g are false. However, in some applications, such as example 1, no very manageable intermediate hypotheses are apparent. The analysis based on the likelihood ratio (15) seems more relevant either when we may assume that either H_f or H_g is true, or when high power is required particularly against the alternative H_g.

A different type of comprehensive model is used by Hotelling [10] in his formulation of the problem of selecting the "best" single predictor. In this approach the hypothesis to be tested is not one of the hypotheses H_f and H_g, but is the hypothesis that in a certain sense H_f and H_g are equally good.

5. Distribution of L_{fg} for simple hypotheses

If both H_f and H_g are simple hypotheses, we may write (15) as

$$L_{fg} = \log \frac{f(\mathbf{Y})}{g(\mathbf{Y})}. \tag{17}$$

For some purposes it may be sufficient to take the observed value of (17) as a measure of the evidence in favor of H_f. The same, incidentally, is not true of (15), since if, for example, $\boldsymbol{\alpha}$ contains more adjustable parameters than $\boldsymbol{\beta}$, it may be expected, other things being equal, to produce a better fit. Two general arguments relating (17) to probabilities of error are available.

First, Barnard [2] and Roy [20] have pointed out that the remark on which Wald's approximations in sequential analysis are based [23] gives a simple inequality in the present application. Suppose for simplicity that we fix a critical value $a > 1$ and regard the data as decisive in favor of H_f if $r_{fg} > a$. Then

$$\frac{P\{R_{fg} > a|H_f\}}{P\{R_{fg} > a|H_g\}} > a, \tag{18}$$

so that in particular

$$P\{R_{fg} > a|H_g\} < \frac{1}{a}, \tag{19}$$

giving an inequality for the probability of error. However in nonsequential problems (19) is usually a poor result because of the inclusion in the critical region of points for which $r_{fg} \gg a$.

Secondly, when the components of Y are independent, L_{fg} is the sum of n independent terms, and we can usually apply the central limit theorem to prove the asymptotic normality of L_{fg}, and hence can obtain approximations to the percentage points of L_{fg} both under H_f and under H_g. The integrals defining the expectations of L_{fg}, namely

$$\int \left[\log \frac{f(y)}{g(y)}\right] f(y)\, dy, \qquad \int \left[\log \frac{f(y)}{g(y)}\right] g(y)\, dy, \tag{20}$$

have been studied in connection with information measures in statistics [13]; the first integral in (20) exceeds the second unless $f(y) \equiv g(y)$.

If the method of section 4 is used for this problem it is natural to consider the family of p.d.f.

$$\frac{[f(y)]^{\lambda}[g(y)]^{1-\lambda}}{\int_{-\infty}^{\infty} [f(y)]^{\lambda}[g(y)]^{1-\lambda}\, dy}, \tag{21}$$

regarding λ as an unknown parameter. The statistic L_{fg} of (17) is sufficient for λ. Cox and Brandwood [7] have applied this family to arrange in order some works of Plato.

6. Distributional problems for composite hypotheses

We discuss now the distribution of L_{fg} under H_f when the hypotheses H_f and H_g are composite. First it is often in principle possible to obtain an exact test, that is, a test of constant size for all $\boldsymbol{\alpha}$. This is so when there are nontrivial sufficient statistics for $\boldsymbol{\alpha}$; the distribution of L_{fg} conditionally on these sufficient statistics is independent of $\boldsymbol{\alpha}$ and can be used to determine a critical region. The method fails only if L_{fg} is a function of the sufficient statistic itself. Most of the problems we have in mind are, however, rather complicated ones for which the conditional distribution is difficult to determine, and therefore it is natural to examine large-sample approximations.

As a preliminary it is necessary to consider the distribution of the maximum likelihood estimator $\hat{\beta}$ under the hypothesis H_f. That is, we consider a maximum likelihood estimator obtained assuming one distribution law to be true, and examine its distribution under some other distribution law.

7. Remarks on the distribution of maximum likelihood estimators

Suppose first that the parameters are one-dimensional and that $Y_1, \cdots, Y_n$ are independently identically distributed with p.d.f. $f(y, \alpha)$ under H_f and $g(y, \beta)$ under H_g. [No confusion will arise from denoting the joint p.d.f. by $f(\mathbf{y}, \alpha)$ and the p.d.f. of a single component by $f(y, \alpha)$]. The log-likelihood for estimating β is

$$L_g(\beta) = \sum \log g(Y_i, \beta) \tag{22}$$

and the maximum likelihood equation is

$$\frac{\partial L_g(\hat{\beta})}{\partial \beta} = 0. \tag{23}$$

Assume for the moment that for given α, $\hat{\beta}$ converges in probability as $n \rightarrow \infty$ to a limit β_α, say. Expanding (23) in the usual way, we have that

$$\hat{\beta} - \beta_\alpha \sim -\frac{\dfrac{\partial L_g(\beta_\alpha)}{\partial \beta}}{\dfrac{\partial^2 L_g(\beta_\alpha)}{\partial \beta^2}}, \tag{24}$$

where numerator and denominator are separately sums of n independent terms. Since $\hat{\beta}$ converges to β_α, we have that

$$E_\alpha\left[\frac{\partial \log g(Y, \beta_\alpha)}{\partial \beta}\right] = 0, \tag{25}$$

where E_α denotes expectation under the p.d.f. $f(y, \alpha)$. Equation (25) can be taken as a definition of β_α. It follows from the form of (24) that $\hat{\beta}$ is asymptotically normal with mean β_α and variance $v_\alpha(\hat{\beta})/n$, where

$$v_\alpha(\hat{\beta}) = \frac{E_\alpha(G_\beta^2)}{[E_\alpha(G_{\beta\beta})]^2}. \tag{26}$$

We have written

$$G_\beta = \frac{\partial \log g(Y, \beta_\alpha)}{\partial \beta}, \qquad G_{\beta\beta} = \frac{\partial^2 \log g(Y, \beta_\alpha)}{\partial \beta^2}, \tag{27}$$

and have assumed the right side of (26) to be finite and nonzero.

If we differentiate (25) with respect to α, we have that, in a notation analogous to (27),

$$E_\alpha(F_\alpha G_\beta) = \frac{d\beta_\alpha}{d\alpha} E_\alpha(-G_{\beta\beta}). \tag{28}$$

When $f \equiv g$, so that $\beta_\alpha \equiv \alpha$, equation (28) is a familiar identity in maximum likelihood theory.

In many applications, especially where an explicit expression for $\hat\beta$ is available, it will be easy to see whether β_α exists. Thus suppose that $f(y, \alpha)$ is the normal distribution of mean α and unit variance, and that $g(y, \beta)$ is the Cauchy distribution centered at β and with unit scale parameter. Then β_α exists and equals α, whereas α_β does not exist, because $\hat\alpha$ is the sample mean and does not converge in probability under the Cauchy distribution.

A simple extension of the argument leading to (13) shows that the joint distribution of $\hat\alpha$ and $\hat\beta$ is asymptotically bivariate normal, and that the covariance of α and β is

$$\frac{d\beta_\alpha/d\alpha}{nE_\alpha(-F_{\alpha\alpha})}, \tag{29}$$

where $F_{\alpha\alpha} = \partial^2 \log f(Y, \alpha)/\partial\alpha^2$. Thus the asymptotic regression coefficient of $\hat\beta$ on $\hat\alpha$, under H_f, is $d\beta_\alpha/d\alpha$, a result that could have been anticipated from the asymptotic sufficiency of $\hat\alpha$.

To generalize the above results to apply to vector parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we assume that under H_f, $\hat{\boldsymbol{\beta}}$ converges in probability, say to $\boldsymbol{\beta}_{\boldsymbol{\alpha}}$. It is convenient to introduce some special notation. Let

$$\begin{aligned} F_i &= \frac{\partial \log f(Y, \boldsymbol{\alpha})}{\partial\alpha_i}, & i &= 1, \cdots, d_f,\\ G_i &= \frac{\partial \log f(Y, \boldsymbol{\beta}_{\boldsymbol{\alpha}})}{\partial\beta_i}, & i &= 1, \cdots, d_g,\\ \mathfrak{F}_i &= \frac{1}{n}\sum_{j=1}^{n} \frac{\partial \log f(Y_j, \boldsymbol{\alpha})}{\partial\alpha_i}, & \mathcal{G}_i &= \frac{1}{n}\sum_{j=1}^{n} \frac{\partial \log g(Y_j, \boldsymbol{\beta}_{\boldsymbol{\alpha}})}{\partial\beta_i},\\ F_{ij} &= \frac{\partial^2 \log f(Y, \boldsymbol{\alpha})}{\partial\alpha_i\partial\alpha_j}, & G_{ij} &= \frac{\partial^2 \log g(Y, \boldsymbol{\beta}_{\boldsymbol{\alpha}})}{\partial\beta_i\partial\beta_j}. \end{aligned} \tag{30}$$

Further let $\{E_{\boldsymbol{\alpha}}(F\cdot\cdot)\}^{ij}$ be the (i, j)th element of the matrix inverse to $E_{\boldsymbol{\alpha}}(F_{ij})$. Define similarly $\{E_{\boldsymbol{\alpha}}(G\cdot\cdot)\}^{ij}$. Note that in all our calculations under H_f, $\boldsymbol{\beta}$ is taken to be $\boldsymbol{\beta}_{\boldsymbol{\alpha}}$. We use the summation convention for repeated suffices.

Corresponding to (24) we have that asymptotically

$$\mathfrak{F}_i + E_{\boldsymbol{\alpha}}(F_{ij})(\hat\alpha_j - \alpha_j) = 0, \tag{31}$$

$$\mathcal{G}_i + E_{\boldsymbol{\alpha}}(G_{ij})(\hat\beta_j - \beta_{\boldsymbol{\alpha},j}) = 0. \tag{32}$$

For the convergence of $\hat{\boldsymbol{\beta}}$, we need $E_{\boldsymbol{\alpha}}(\mathcal{G}_i) = 0$, that is,

$$E_{\boldsymbol{\alpha}}(G_i) = 0. \tag{33}$$

Equations (31) and (32) give

$$\hat\alpha_i - \alpha_i = -\{E_{\boldsymbol{\alpha}}(F\cdot\cdot)\}^{ij}\,\mathfrak{F}_j, \tag{34}$$

$$\hat{\beta}_i - \beta_{\alpha,i} = -\{E_\alpha(G\cdot\cdot)\}^{ij}\,\mathcal{G}_j. \tag{35}$$

Now $\mathcal{F}_j$ is the mean of n independent identically distributed random variables of mean 0, and so also is $\mathcal{G}_j$. Therefore

$$E_\alpha(\mathcal{F}_j\mathcal{F}_l) = \frac{1}{n}E_\alpha(F_jF_l), \tag{36}$$

$$E_\alpha(\mathcal{F}_j\mathcal{G}_l) = \frac{1}{n}E_\alpha(F_jG_l), \tag{37}$$

$$E_\alpha(\mathcal{G}_j\mathcal{G}_l) = \frac{1}{n}E_\alpha(G_jG_l). \tag{38}$$

The asymptotic covariance matrix of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ follows immediately from (34) to (38). To simplify the answer we use the standard result that

$$E_\alpha(F_jF_l) + E_\alpha(F_{jl}) = 0. \tag{39}$$

Further, a result analogous to (28) is obtained by differentiating (33) with respect to α_j. This gives

$$E_\alpha(G_{il})\frac{\partial\beta_{\alpha,l}}{\partial\alpha_j} + E_\alpha(G_iF_j) = 0. \tag{40}$$

We have now from (34), (36), and (39), the standard result that

$$\mathrm{Cov}\,(\hat{\alpha}_i - \alpha_i, \hat{\alpha}_j - \alpha_j) = -\frac{1}{n}\{E_\alpha(F\cdot\cdot)\}^{ij}, \tag{41}$$

and from (34), (35), (37), and (40),

$$\mathrm{Cov}\,(\hat{\alpha}_i - \alpha_i, \hat{\beta}_j - \beta_{\alpha,j}) = -\frac{1}{n}\{E_\alpha(F\cdot\cdot)\}^{il}\frac{\partial\beta_{\alpha,j}}{\partial\alpha_l}, \tag{42}$$

and

$$\mathrm{Cov}\,(\hat{\beta}_i - \beta_{\alpha,i}, \hat{\beta}_j - \beta_{\alpha,j}) = \frac{1}{n}\{E_\alpha(G\cdot\cdot)\}^{il}E_\alpha(G_lG_m)\{E_\alpha(G\cdot\cdot)\}^{mj}. \tag{43}$$

To sum up, $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is asymptotically normal with mean $(\boldsymbol{\alpha}, \boldsymbol{\beta}_\alpha)$ and covariance matrix given by (41) to (43). The results generalize immediately when the Y_i are distributed independently, but not identically.

For mathematical analysis of general problems in which the number of components in $\boldsymbol{\alpha}$ is not less than that in $\boldsymbol{\beta}$, it would frequently be convenient to parametrize H_f in such a way that $\beta_{\alpha,i} = \alpha_i$. We shall, however, not do this here.

In simple cases it may be possible, in work with data, to evaluate the expectations in (41) to (43) mathematically and then to replace $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\alpha}}$. Otherwise it is permissible, as in ordinary maximum likelihood estimation, to replace the expectation by a consistent estimate of it. Thus

$$E_\alpha(-F_{\alpha\alpha}) = E_\alpha\left[-\frac{\partial^2 \log f(Y, \hat{\alpha})}{\partial\alpha^2}\right] \tag{44}$$

can be replaced by

$$-\frac{1}{n}\sum\frac{\partial^2 \log f(y_i, \hat{\alpha})}{\partial\alpha^2}, \tag{45}$$

and so on. That is, assuming H_f to be true, all the quantities occurring in the above formulas can be estimated consistently from the data, and this is enough for what follows.

8. Tests based on L_{fg}

Having computed l_{fg}, how do we interpret it? First, a large positive value is evidence against H_g, a large negative value evidence against H_f. The formal proof of this follows on writing

$$
\begin{aligned}
(46)\qquad L_{fg} &= L_f(\hat{\boldsymbol{\alpha}}) - L_g(\hat{\boldsymbol{\beta}}) \\
&= \{L_f(\boldsymbol{\alpha}) - L_g(\boldsymbol{\beta}_\alpha)\} + \{L_f(\hat{\boldsymbol{\alpha}}) - L_f(\boldsymbol{\alpha})\} - \{L_g(\hat{\boldsymbol{\beta}}) - L_g(\boldsymbol{\beta}_\alpha)\}.
\end{aligned}
$$

Expansion of the last two terms shows them to be of order one in probability, whereas under H_f the first term has a positive expectation of order n and standard deviation of order $\sqrt{n}$.

This might suggest trying to examine whether l_{fg} is in some sense significantly positive or significantly negative. This sort of approach would, however, be appropriate when the hypothesis under test is not H_f, but is the hypothesis that H_f and H_g are "equally good"; see specially the discussion in [25]. Note also that, even from the Bayesian point of view, and even if H_f and H_g have equal prior probability, the critical value of zero for l_{fg}, that is, of one for r_{fg}, has justification only under certain conditions on the prior probability densities $p_f(\boldsymbol{\alpha})$, $p_g(\boldsymbol{\beta})$. In fact in the asymptotic equation (14), the final factor is unity only if

$$
(47)\qquad (2\pi)^{d_f/2} p_f(\boldsymbol{\alpha})\, \Delta_\alpha^{-1/2} = (2\pi)^{d_g/2} p_g(\boldsymbol{\beta})\, \Delta_\beta^{-1/2}.
$$

This is so if $p_f(\boldsymbol{\alpha})$ and $p_g(\boldsymbol{\beta})$ assign equal prior probabilities to standard confidence ellipsoids based on $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, and it is hard to see a general reason for expecting this.

If we take H_f as the hypothesis under test and H_g as the alternative, it seems reasonable to consider for a test statistic

$$
(48)\qquad T_f = L_{fg} - E_{\hat{\alpha}}(L_{fg}),
$$

that is, to compare L_{fg} with the best estimate of the value we would expect it to take under H_f. A large negative value of T_f would lead to the "rejection" of H_f.

9. An example

To examine (48) more closely, consider example 1. Here $H_f^{(1)}$ assigns each Y_i the p.d.f.

$$
(49)\qquad f(y, \boldsymbol{\alpha}) = \frac{1}{y(2\pi\alpha_2)^{1/2}} \exp\left[-\frac{(\log y - \alpha_1)^2}{2\alpha_2}\right],
$$

whereas for $H_g^{(1)}$ the p.d.f. is

$$g(y, \beta) = \beta^{-1}e^{-y/\beta}. \tag{50}$$

Now $\hat{\beta}$ is the sample mean of the Y and therefore converges under $H_f^{(1)}$ to the mean of the distribution (49), that is, to $\exp(\alpha_1 + \alpha_2/2)$. Thus

$$\beta_{\alpha} = \exp\left[\alpha_1 + \frac{1}{2}\alpha_2\right]. \tag{51}$$

Further, noting that $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are the sample mean and variance of the log Y, we find that

$$L_{fg} = -\frac{1}{2}n \log 2\pi\hat{\alpha}_2 + \frac{1}{2}n - n\hat{\alpha}_1 + n \log \hat{\beta}. \tag{52}$$

Incidentally, a practical objection to the criterion (52) in some situations is that it depends rather critically on observations with very small values, and such observations are liable to be subject to relatively large recording errors.

We have that

$$\log \frac{f(y, \boldsymbol{\alpha})}{g(y, \beta_{\boldsymbol{\alpha}})} = -\frac{1}{2}\log(2\pi\alpha_2) - \frac{(\log y - \alpha_1)^2}{2\alpha_2} - \log y + \alpha_1 + \frac{1}{2}\alpha_2 + y \exp\left(\alpha_1 + \frac{1}{2}\alpha_2\right), \tag{53}$$

and therefore

$$E_{\boldsymbol{\alpha}}\left[\log \frac{f(Y, \boldsymbol{\alpha})}{g(Y, \beta_{\boldsymbol{\alpha}})}\right] = -\frac{1}{2}\log(2\pi\alpha_2) + \frac{1}{2} + \frac{1}{2}\alpha_2, \tag{54}$$

so that

$$T_f = n \log \frac{\hat{\beta}}{\beta_{\hat{\boldsymbol{\alpha}}}}, \tag{55}$$

because T_f is given by L_{fg} minus n times expression (54).

An equivalent test statistic is

$$\frac{T_f}{n} = \log \frac{\hat{\beta}}{\beta_{\hat{\boldsymbol{\alpha}}}}. \tag{56}$$

From the results established in section 7 about the asymptotic distribution of $(\hat{\boldsymbol{\alpha}}, \hat{\beta})$, it follows immediately that (56) is asymptotically normal with mean zero. To calculate the asymptotic variance of (56), we need first $\mathbf{C}_{\boldsymbol{\alpha}}$, the asymptotic covariance matrix of $(\hat{\boldsymbol{\alpha}}, \hat{\beta})$; this can be calculated, either directly or from (41) to (43), to be given by

$$n\mathbf{C}_{\boldsymbol{\alpha}} = \begin{pmatrix} \alpha_2 & 0 & \alpha_2 e^{\beta_{\boldsymbol{\alpha}}} \\ 0 & 2\alpha_2^2 & \alpha_2^2 e^{\beta_{\boldsymbol{\alpha}}} \\ \alpha_2 e^{\beta_{\boldsymbol{\alpha}}} & \alpha_2^2 e^{\beta_{\boldsymbol{\alpha}}} & \beta_{\boldsymbol{\alpha}}^2(e^{\alpha_2} - 1) \end{pmatrix}. \tag{57}$$

Now (56) can be written asymptotically

$$(58)\qquad \log\left(1+\frac{\hat{\beta}-\beta_\alpha}{\beta_\alpha}\right)-\log\left(1+\frac{\beta_{\hat{\alpha}}-\beta_\alpha}{\beta_\alpha}\right)$$

$$\sim\frac{\hat{\beta}-\beta_\alpha}{\beta_\alpha}-\frac{1}{\beta_\alpha}\left[\frac{\partial\beta_\alpha}{\partial\alpha_1}(\hat{\alpha}_1-\alpha_1)+\frac{\partial\beta_\alpha}{\partial\alpha_2}(\hat{\alpha}_2-\alpha_2)\right]$$

$$=\left(-1,-\frac{1}{2},\frac{1}{\beta_\alpha}\right)(\hat{\alpha}_1-\alpha_1,\hat{\alpha}_2-\alpha_2,\hat{\beta}-\beta_\alpha)'$$

$$=l_\alpha(\hat{\alpha}_1-\alpha_1,\hat{\alpha}_2-\alpha_2,\hat{\beta}-\beta_\alpha)',$$

say. Therefore the asymptotic variance of (56) is $\mathbf{l}_\alpha\mathbf{C}_\alpha\mathbf{l}'_\alpha$ and simple calculation shows this to be

$$(59)\qquad \frac{1}{n}\left(e^{\alpha_2}-1-\alpha_2-\frac{1}{2}\alpha_2^2\right),$$

which is estimated consistently by replacing α_2 by $\hat{\alpha}_2$.

Thus, finally, we test $H_f^{(1)}$ by referring

$$(60)\qquad \frac{\log\dfrac{\hat{\beta}}{\beta_{\hat{\alpha}}}}{\left[\dfrac{1}{n}\left(e^{\hat{\alpha}_2}-1-\hat{\alpha}_2-\dfrac{1}{2}\hat{\alpha}_2^2\right)\right]^{1/2}}$$

to the tables of the standard normal distribution, large negative values of (60) being evidence of departures from $H_f^{(1)}$ in the direction of $H_g^{(1)}$. Clearly there are a number of asymptotically equivalent forms for the numerator of (60).

Suppose now that $H_g^{(1)}$ is true. The statistics $\hat{\alpha}_1$ and $\hat{\alpha}_2$ converge under $H_g^{(1)}$ to the mean and variance of $\log Y$, where Y has the exponential distribution (50). Therefore

$$(61)\qquad \alpha_{1,\beta}=\log\beta+\psi(1),$$

$$(62)\qquad \alpha_{2,\beta}=\psi'(1),$$

where, in the usual notation, $\psi(x)=d\log\Gamma(x)/dx$; $\psi(1)=-0.5772$, and $\psi'(1)=1.6449$.

It follows from (61) and (62) that under $H_g^{(1)}$ the numerator of (60) converges to

$$(63)\qquad \log\frac{\beta}{\exp\left(\alpha_{1,\beta}+\frac{1}{2}\alpha_{2,\beta}\right)}=-\psi(1)-\frac{1}{2}\psi'(1).$$

Since this is different from zero, the consistency of the test based on T_f is proved. Asymptotic power properties can be obtained by calculating also the variance of $\hat{\beta}-\beta_{\hat{\alpha}}$ under $H_g^{(1)}$; such calculations would be relevant only for values of $\boldsymbol{\alpha}$ for which a large n is required in order to attain moderate power.

If we take $H_g^{(1)}$ as the hypothesis under test, we need to calculate

$$(64)\qquad T_g=L_{fg}-E_\beta(L_{fg}).$$

Equation (52) still gives L_{fg}, but (53) must be replaced by

$$\log \frac{f(y, \alpha_\beta)}{g(y, \beta)} = -\frac{(\log y - \alpha_{1,\beta})^2}{2\alpha_{2,\beta}} - \log y - \frac{1}{2}\log(2\pi\alpha_{2,\beta}) + \log\beta + \frac{y}{\beta}. \tag{65}$$

We obtain, from (52), (61), (62) and (65), the result that

$$T_g = -n(\hat{\alpha}_1 - \alpha_{1,\beta}) - \frac{1}{2} n \log \frac{\hat{\alpha}_2}{\psi'(1)}, \tag{66}$$

where $\psi'(1)$ is, in virtue of (62), the constant value of $\alpha_{2,\beta}$. An equivalent test statistic is

$$-(\hat{\alpha}_1 - \alpha_{1,\beta}) - \frac{1}{2}\log\frac{\hat{\alpha}_2}{\psi'(1)}. \tag{67}$$

Under $H_g^{(1)}$, (67) has an asymptotic normal distribution of mean zero and a variance calculated by replacing (67) by

$$\begin{aligned} &-(\hat{\alpha}_1 - \alpha_{1,\beta}) - \frac{\hat{\alpha}_2 - \psi'(1)}{2\psi'(1)} + \frac{1}{\beta}(\hat{\beta} - \beta) \\ &\qquad \sim \left(-1, \frac{-1}{2\psi'(1)}, \frac{1}{\beta}\right)(\hat{\alpha}_1 - \alpha_{1,\beta}, \hat{\alpha}_2 - \alpha_{2,\beta}, \hat{\beta} - \beta)' \\ &\qquad = \mathbf{m}_\beta(\hat{\alpha}_1 - \alpha_{1,\beta}, \hat{\alpha}_2 - \alpha_{2,\beta}, \hat{\beta} - \beta)', \end{aligned} \tag{68}$$

say. The asymptotic covariance matrix $\mathbf{C}_\beta$ of $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta})$ is found either from the general formulas of section 7, or from first principles, to be given by

$$n\mathbf{C}_\beta = \begin{pmatrix} \psi'(1) & \psi''(1) & \beta \\ \psi''(1) & \psi'''(1) + 2[\psi'(1)]^2 & 0 \\ \beta & 0 & \beta^2 \end{pmatrix}. \tag{69}$$

The asymptotic variance of (67) is thus

$$\mathbf{m}_\beta \mathbf{C}_\beta \mathbf{m}_\beta' = \frac{1}{n}\left[\psi'(1) - \frac{1}{2} + \frac{\psi''(1)}{\psi'(1)} + \frac{\psi'''(1)}{4[\psi'(1)]^2}\right] = \frac{0.856}{n}. \tag{70}$$

Therefore the test of $H_g^{(1)}$ is to refer

$$\frac{-(\hat{\alpha}_1 - \alpha_{1,\beta}) - \frac{1}{2}\log\frac{\hat{\alpha}_2}{\psi'(1)}}{0.925/\sqrt{n}} \tag{71}$$

to the standard normal tables, large positive values of (71) being evidence of a departure from $H_g^{(1)}$ in the direction of $H_f^{(1)}$.

Illustrative example. A random sample of size 50 from the exponential distribution of unit mean gave for the mean and variance of $\log y$, the values $\hat{\alpha}_1 = -0.6404$, $\hat{\alpha}_2 = 1.2553$, and for the mean of y, the value $\hat{\beta} = 0.8102$. The log likelihood-ratio l_{fg} is -5.15, corresponding to $r_{fg} = 1/170$, apparently a large factor in favor of $H_g^{(1)}$, the exponential distribution. However, we cannot interpret this value in a direct way, because we do not know the other factors in (14).

If we take $H_f^{(1)}$ as the hypothesis to be tested we get from (56) the ratio $T_f/n = -0.1977$, with an estimated standard error of 0.0965, leading, by (60), to the critical ratio 2.048. There is therefore quite strong evidence of a departure from the log-normal family in the direction of the exponential. If we now take $H_g^{(1)}$ as the hypothesis to be tested we get from (67), $T_g/n = -0.0121$, with an asymptotic standard error 0.131. The critical ratio (71) is thus 0.092, indicating good agreement with the exponential family, in the respect tested. A calculation of power and a detailed comparison with other tests have not been attempted.

10. Some results for the Koopman-Darmois family

In the special case discussed in section 9, the test of $H_f^{(1)}$ reduces to an examination of whether $\hat{\beta} - \beta_{\hat{\alpha}}$ differs significantly from zero, and the test of $H_g^{(1)}$ reduces to an examination of whether a certain asymptotically linear combination of $(\hat{\alpha}_1 - \alpha_{1,\hat{\beta}}, \hat{\alpha}_2 - \alpha_{2,\hat{\beta}})$ differs significantly from zero. This suggests the question: when does the test of H_f reduce to an examination of whether an asymptotically linear combination of the components of $\boldsymbol{\beta}_{\hat{\boldsymbol{\alpha}}} - \hat{\boldsymbol{\beta}}$ differs significantly from zero? When the test does have this form, the asymptotic normal distribution of the test criterion has mean zero and a variance determined by the theory of section 7, and hence in such cases our problem is in principle solved.

Obviously the likelihood ratio test cannot take this form always. For example, if the hypotheses H_f and H_g correspond respectively to a Poisson distribution and to a geometric distribution, α and β may both be taken as the distribution mean. Since the sample mean is a sufficient statistic under both H_f and H_g, $\beta_\alpha = \alpha$, $\alpha_\beta = \beta$, and $\alpha = \beta = \beta_\alpha = \alpha_\beta$, so that the examination of $\beta_{\hat{\alpha}} - \hat{\beta}$ leads to a vacuous result.

To try to generalize the results of section 9, it is natural to turn to the so-called Koopman-Darmois family. Since most of the problems which we want to answer do not concern random samples, we shall not restrict the Y_i to be identically distributed.

Let $Y_1, \cdots, Y_n$ be independently distributed with p.d.f. under H_f

$$\exp\left[\sum_j A_{ij}^{(f)}(\alpha_j^*)B_{ij}^{(f)}(y) + C_i^{(f)}(y) + D_i^{(f)}(\boldsymbol{\alpha})\right], \tag{72}$$

and under H_g

$$\exp\left[\sum_j A_{ij}^{(g)}(\beta_j^*)B_{ij}^{(g)}(y) + C_i^{(g)}(y) + D_i^{(g)}(\boldsymbol{\beta})\right], \tag{73}$$

where the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are split into components α_j^*, β_j^* in such a way that each term in the leading summations refers to a single component only.

As an instance of (72), consider the linear regression model of example 2. Here the p.d.f. of Y_i under the hypothesis $H_f^{(2)}$

$$\begin{aligned}\frac{1}{(2\pi\alpha_0)^{1/2}} &\exp\left[-\frac{(y-\alpha_1-\alpha_2 x_i)^2}{2\alpha_0}\right] \\ &= \exp\left[-\frac{y^2}{2\alpha_0} + \frac{y\alpha_1}{\alpha_0} + \frac{y\alpha_2 x_i}{\alpha_0} - \frac{1}{2}\log(2\pi\alpha_0) - \frac{(\alpha_1+\alpha_2 x_i)^2}{2\alpha_0}\right].\end{aligned} \tag{74}$$

This is of the form (56) with $C_i^{(f)}(y) \equiv 0$, provided that we take $\alpha_1^* = 1/\alpha_0$, $\alpha_2^* = \alpha_1/\alpha_0$, $\alpha_3^* = \alpha_2/\alpha_0$.

It is now straightforward, using (72) and (73), to compute the test statistic T_f. For

$$L_f(\boldsymbol{\alpha}) = \sum_{i,j} A_{ij}^{(f)}(\alpha_j^*)B_{ij}^{(f)}(y_i) + \sum_i C_i^{(f)}(y_i) + \sum_i D_i^{(f)}(\boldsymbol{\alpha}) \tag{75}$$

and the corresponding expectation is the sum over i of the logarithms of expressions like (72), and so is

$$\sum_{i,j} A_{ij}^{(f)}(\alpha_j^*)E_{\boldsymbol{\alpha}}[B_{ij}^{(f)}(Y_i)] + \sum_i E_{\boldsymbol{\alpha}}[C_i^{(f)}(Y_i)] + \sum_i D_i^{(f)}(\boldsymbol{\alpha}). \tag{76}$$

The maximum likelihood equation for $\hat{\alpha}_j^*$ is

$$\sum_i \frac{dA_{ij}^{(f)}(\hat{\alpha}_j^*)}{d\alpha_j^*} B_{ij}^{(f)}(y_i) + \sum_i \frac{\partial}{\partial \alpha_j} D_i^{(f)}(\hat{\boldsymbol{\alpha}}) = 0. \tag{77}$$

Also the expectation of the logarithmic derivative of (72) is zero,

$$\frac{dA_{ij}^{(f)}(\alpha_j^*)}{d\alpha_j^*} E_{\boldsymbol{\alpha}}[B_{ij}^{(f)}(Y_i)] + \frac{\partial}{\partial \alpha_j^*} D_i^{(f)}(\boldsymbol{\alpha}) = 0. \tag{78}$$

Combination of (75) to (78) gives

$$L_f(\hat{\boldsymbol{\alpha}}) - E_{\hat{\boldsymbol{\alpha}}}[L_f(\hat{\boldsymbol{\alpha}})] = \sum_i \{C_i^{(f)}(y_i) - E_{\hat{\boldsymbol{\alpha}}}[C_i^{(f)}(Y)]\}. \tag{79}$$

The corresponding expression arising from the likelihood under H_g is obtained by a similar, but slightly more complicated, argument; the answer is

$$L_g(\hat{\boldsymbol{\beta}}) - E_{\hat{\boldsymbol{\alpha}}}[L_g(\boldsymbol{\beta}_{\hat{\boldsymbol{\alpha}}})] = \sum_i \{C_i^{(g)}(y_i) - E_{\hat{\boldsymbol{\alpha}}}[C_i^{(g)}(Y_i)]\} + H_g(\boldsymbol{\beta}_{\hat{\boldsymbol{\alpha}}}) - H_g(\hat{\boldsymbol{\beta}}), \tag{80}$$

where

$$H_g(\boldsymbol{\beta}) = \sum_{i,j} \frac{\frac{\partial}{\partial \beta_j^*}[D_i^{(g)}(\boldsymbol{\beta})A_{ij}(\beta_j^*)]}{\frac{\partial}{\partial \beta_j^*}[A_{ij}(\beta_j^*)]}. \tag{81}$$

The test statistic T_f, equal to $L_{fg} - E_{\hat{\boldsymbol{\alpha}}}[L_{fg}]$, is the difference between (79) and (80).

It follows that a simple sufficient condition that the T_f test reduces to a comparison of $\hat{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}_{\hat{\boldsymbol{\alpha}}}$ is that in the representations (72) and (73)

$$C_i^{(f)}(y) \equiv C_i^{(g)}(y). \tag{82}$$

In example 1, discussed in section 9, and in the regression model (74), the C functions are identically zero, so that (82) is satisfied; indeed applications in which the C are zero are likely to be the main common case in which (82) holds.

When (82) is satisfied

$$T_f = H_g(\boldsymbol{\beta}_{\hat{\boldsymbol{\alpha}}}) - H_g(\hat{\boldsymbol{\beta}}). \tag{83}$$

Now $\boldsymbol{\beta}_{\hat{\boldsymbol{\alpha}}} - \hat{\boldsymbol{\beta}}$ has an asymptotic multivariate normal distribution with mean zero and covariance matrix that can be found from the results of section 7.

Hence, by applying the Δ-method of asymptotic standard errors to (83), the variance of T_f under H_f can be consistently estimated and the hypothesis H_f tested in a way directly analogous to (60). The general expression for the test criterion is rather complicated and there seems no point in writing it down; the method fails if the limiting normal distribution of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is singular in such a way that the apparent variance of T_f is zero.

The test based on (83) gives an answer in principle to a range of problems of which examples 1 to 4 of section 2 are instances. Clearly, however, a great deal remains to be done on special cases, especially if the adequacy of the approximations is to be assessed.

A case where the results of this section fail badly is provided by taking the hypothesis $H_g^{(4)}$ of example 4 as the hypothesis under test. Here, because negative Y_i may arise under $H_g^{(4)}$, the estimate $\hat{\boldsymbol{\alpha}}$ does not converge in probability under H_g, that is, $\boldsymbol{\alpha}_{\boldsymbol{\beta}}$ does not exist.

11. More general situations

In this section we note briefly that the results of sections 9 and 10 can be extended. First, if the Koopman-Darmois laws (72) and (73) hold, and if condition (82) is not satisfied, the test statistic T_f is given not by (83) but by

$$T_f = H_g(\boldsymbol{\beta}_{\hat{\boldsymbol{\alpha}}}) - H_g(\hat{\boldsymbol{\beta}}) + \sum \{C_i(y_i) - E_{\hat{\boldsymbol{\alpha}}}[C_i(Y)]\}, \tag{84}$$

where

$$C_i(y) = C_i^{(f)}(y) - C_i^{(g)}(y). \tag{85}$$

To compute the asymptotic variance of T_f under the hypothesis H_f, we need the asymptotic covariance matrix not just of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$, but of $[\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \sum C_i(Y_i)/n]$. An expression for this can be found.

More generally, if we do not restrict ourselves to distributions of the Koopman-Darmois type, L_{fg} can be written in the form (46), the last two terms being ignored because they are of order one in probability. Further, considering for simplicity a problem in which α is one-dimensional, we have that

$$E_{\alpha}(L_{fg}) - E_{\alpha}(L_{fg}) \sim (\hat{\alpha} - \alpha) \frac{dE_{\alpha}(L_{fg})}{d\alpha}; \tag{86}$$

the asymptotic normality of $T_f = L_{fg} - E_{\hat{\alpha}}(L_{fg})$ can be proved from this and the retention of terms of order one in probability would give a further approximation.

These results will not be followed up in the present paper.

12. A regression problem

An illustration of the difficulties that can arise in applying the results of section 10 is provided by examples 2 and 3 of section 2, which will now be discussed briefly. Let α_0, β_0 be the residual variances under the linear models

$$H_f: E(\mathbf{Y}) = \mathbf{a}\boldsymbol{\alpha};$$
(87)
$$H_g: E(\mathbf{Y}) = \mathbf{b}\boldsymbol{\beta},$$

of example 3. Then

(88) $$L_{fg} = \frac{1}{2} n \log \frac{\hat{\beta}_0}{\hat{\alpha}_0}.$$

Conditions (72) and (73) are satisfied and it might therefore be expected that the T_f test would be equivalent to a comparison of $\hat{\beta}_0 - E_{\hat{\alpha}}(\hat{\beta}_0)$ with its large-sample standard error, the limiting distribution being normal. However, if we let $n \to \infty$, with the ranks of $\mathbf{a}$ and $\mathbf{b}$ fixed, we find that the limiting distribution is not normal. The reason, roughly speaking, is that the statistics $\hat{\alpha}_0$ and $\hat{\beta}_0$ contain a large common part, the sum of squares residual to $(\mathbf{a} \vdots \mathbf{b})$, and that the test-statistic corresponds to a singularity in the limiting normal distribution of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$.

If the matrices $(\mathbf{a}'\mathbf{a})$, $(\mathbf{b}'\mathbf{b})$ are arranged to be nonsingular, and of size ν_a, ν_b, it is straightforward to show that $\hat{\beta}_0$ minus an estimate of its expectation under H_f can be written in the form

(89) $$\hat{\beta}_0 - \frac{\hat{\boldsymbol{\alpha}}'\mathbf{a}'\mathbf{r}_b\mathbf{a}\hat{\boldsymbol{\alpha}}}{n - \nu_b} - \left\{1 - \operatorname{trace}\left[\mathbf{a}(\mathbf{a}'\mathbf{a})^{-1}(\mathbf{a}'\mathbf{r}_b\mathbf{a})(\mathbf{a}'\mathbf{a})^{-1}\mathbf{a}'\right]\frac{1}{n - \nu_b}\right\}\hat{\alpha}_0,$$

where

(90) $$\mathbf{r}_b = \mathbf{I} - \mathbf{b}(\mathbf{b}'\mathbf{b})^{-1}\mathbf{b}';$$

the statistic (89) has mean exactly zero under H_f.

Note that this statistic is defined in certain situations where the variance ratio for regression on $\mathbf{b}$ adjusting for $\mathbf{a}$ is not defined, namely when rank $(\mathbf{a} \vdots \mathbf{b}) = n$. This situation is uncommon but could arise in the study of factorial experiments with many factors.

It can be shown that when the spaces spanned by the columns of $\mathbf{a}$ and $\mathbf{b}$ are orthogonal, and the rank of $(\mathbf{a} \vdots \mathbf{b})$ is less than n, the statistic (89) is proportional to the difference between the mean square for regression on $\mathbf{b}$ adjusting for $\mathbf{a}$, and the mean square residual to $\mathbf{a}$ and $\mathbf{b}$, so that a test essentially equivalent to the classical F-test is obtained.

Finally consider example 2. We can work either from (89), or from an appropriate canonical form for the hypotheses H_f and H_g to show that (71) reduces to

(91) $$-2r(1 - r^2)^{1/2}\frac{UV}{n - 2} - \frac{r^2(n - 3)}{(n - 2)^2}V^2 + \frac{r^2(n - 3)}{(n - 2)^2}s^2,$$

where r is the sample correlation coefficient between the two independent variables x and $\log x$, s^2 is the mean square about the combined regression relation, and U and V are respectively proportional to the total regression coefficient of Y on x, and to that of Y on $\log x$ for fixed x, both scaled so as to have the variance of a single observation.

Under H_f, (91) has mean zero, whereas under H_g its expectation is negative.

To test H_f we can argue conditionally on U, when the variance of (91) can be estimated in a straightforward way.

It is not clear that (91) has any advantages over the usual one-sided test of V/s; further when there is appreciable regression U will be large, of order $\sqrt{n}$, and (91) is then essentially an examination of the sign of V. It is possible that when there is more than one parameter at stake something essentially different from the conventional test emerges.

13. Discussion

This paper leaves a great deal unsolved, especially

(i) the detailed theory of special cases, in particular the whole matter of the adequacy of the asymptotic theory;

(ii) the treatment of "discontinuous" parameters;

(iii) the calculation of the power functions of the tests proposed here [see, however, the comments following (63)];

(iv) the discussion of possible asymptotic optimum properties. It is known that the ratio of maximum likelihoods can lead to a silly test in some small-sample situations [14], but that in the common parametric cases optimum properties hold asymptotically [22]. Is the same true for the tests using T_f and T_g?

REFERENCES

[1] J. Aitchison and S. D. Silvey, "Maximum likelihood estimation procedures and associated tests of significance," *J. Roy. Statist. Soc., Ser. B*, Vol. 22 (1960), pp. 154–171.

[2] G. A. Barnard, "Review of A. Wald's 'Sequential Analysis,'" *J. Amer. Statist. Assoc.*, Vol. 42 (1957), pp. 658–664.

[3] ———, "Control charts and stochastic processes," *J. Roy. Statist. Soc., Ser. B*, Vol. 21 (1959), pp. 239–271.

[4] H. Chernoff, "On the distribution of the likelihood ratio," *Ann. Math. Statist.*, Vol. 25 (1954), pp. 573–578.

[5] C. L. Chiang, J. L. Hodges, and J. Yerushalmy, "Statistical problems in medical diagnosis," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1956, Vol. 4, pp. 121–133.

[6] C. P. Cox, "The analysis of Latin square designs with individual curvatures in one direction," *J. Roy. Statist. Soc., Ser. B*, Vol. 20 (1958), pp. 193–204.

[7] D. R. Cox and L. Brandwood, "On a discriminatory problem connected with the works of Plato," *J. Roy. Statist. Soc., Ser. B*, Vol. 21 (1959), pp. 195–200.

[8] D. A. Darling and M. Kac, "On occupations times for Markoff processes," *Trans. Amer. Math. Soc.*, Vol. 84 (1957), pp. 444–458.

[9] M. J. R. Healy, "A significance test for the difference in efficiency between two predictors," *J. Roy. Statist. Soc., Ser. B*, Vol. 17 (1955), pp. 266–268.

[10] H. Hotelling, "The selection of variates for use in prediction with some comments on the general problem of nuisance parameters," *Ann. Math. Statist.*, Vol. 11 (1940), pp. 271–283.

[11] J. O. Irwin, "The distribution of the logarithm of survival times when the true law is exponential," *J. Hygiene*, Vol. 42 (1942), pp. 328–333.

[12] H. JEFFREYS, *Theory of Probability*, Oxford, Oxford University Press, 1950 (2nd ed.).

[13] S. KULLBACK, *Information Theory and Statistics*, New York, Wiley, 1959.

[14] E. L. LEHMANN, "Some principles of the theory of testing hypotheses," *Ann. Math. Statist.*, Vol. 21 (1950), pp. 1–26.

[15] D. V. LINDLEY, "The use of prior probability distributions in statistical inference," *Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1961, Vol. I, pp. 453–468.

[16] E. S. PAGE, "A test for a change in a parameter occurring at an unknown point," *Biometrika*, Vol. 42 (1955), pp. 523–527.

[17] ———, "On problems in which a change in parameter occurs at an unknown point," *Biometrika*, Vol. 44 (1957), pp. 248–252.

[18] R. E. QUANDT, "The estimation of the parameter of a linear regression system obeying two separate regimes," *J. Amer. Statist. Assoc.*, Vol. 53 (1958), pp. 873–880.

[19] A. S. C. ROSS, "Philological probability problems," *J. Roy. Statist. Soc., Ser. B*, Vol. 12 (1950), pp. 19–40.

[20] A. D. ROY, "A further statistical note on the distribution of individual output," *Economic J.*, Vol. 60 (1950), pp. 831–836.

[21] S. D. SILVEY, "The Lindisfarne scribe's problem," *J. Roy. Statist. Soc., Ser. B*, Vol. 20 (1958), pp. 93–101.

[22] A. WALD, "Tests of statistical hypotheses concerning several parameters when the number of parameters is large," *Trans. Amer. Math. Soc.*, Vol. 54 (1943), pp. 426–482.

[23] ———, *Sequential Analysis*, New York, Wiley, 1947.

[24] E. J. WILLIAMS, "The comparison of regression variables," *J. Roy. Statist. Soc., Ser. B*, Vol. 21 (1959), pp. 396–399.

[25] ———, *Regression Analysis*, New York, Wiley, 1959, chapter 5.