ISSN 1440-771X

MONASH University

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Dual P-Values, Evidential Tension and Balanced Tests

D. S. Poskitt and Arivalzahan Sengarapillai

June 2010

Working Paper 15/10

Dual P-Values, Evidential Tension and Balanced Tests

D. S. Poskitt
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: Donald.Poskitt@buseco.monash.edu.au

Arivalzahan Sengarapillai
Department of Mathematics and Statistics,
University of Jaffna,
Sri Lanka.

23 June 2010

JEL classification: C12, C44, C52

Dual P-Values, Evidential Tension and Balanced Tests

Abstract:

In the classical approach to statistical hypothesis testing the role of the null hypothesis $\mathscr{H}_0$ and the alternative $\mathscr{H}_1$ is very asymmetric. Power, calculated from the distribution of the test statistic under $\mathscr{H}_1$, is treated as a theoretical construct that can be used to guide the choice of an appropriate test statistic or sample size, but power calculations do not explicitly enter the testing process in practice. In a significance test a decision to accept or reject $\mathscr{H}_0$ is driven solely by an examination of the strength of evidence against $\mathscr{H}_0$, summarized in the P-value calculated from the distribution of the test statistic under $\mathscr{H}_0$. A small P–value is taken to represent strong evidence against $\mathscr{H}_0$, but it need not necessarily indicate strong evidence in favour of $\mathscr{H}_1$. More recently, Moerkerke et al. (2006) have suggested that the special status of $\mathscr{H}_0$ is often unwarranted or inappropriate, and argue that evidence against $\mathscr{H}_1$ can be equally meaningful. They propose a *balanced* treatment of both $\mathscr{H}_0$ and $\mathscr{H}_1$ in which the classical P–value is supplemented by the P–value derived under $\mathscr{H}_1$. The alternative P–value is the dual of the null P–value and summarizes the evidence against a target alternative. Here we review how the dual P–values are used to assess the evidential tension between $\mathscr{H}_0$ and $\mathscr{H}_1$, and use decision theoretic arguments to explore a *balanced* hypothesis testing technique that exploits this evidential tension. The operational characteristics of *balanced* hypothesis tests is outlined and their relationship to conventional notions of optimal tests is laid bare. The use of *balanced* hypothesis tests as a conceptual tool is illustrated via model selection in linear regression and their practical implementation is demonstrated by application to the detection of cancer-specific protein markers in mass spectroscopy.

Keywords: *balanced* test, P- value, dual P–values, evidential tension, null hypothesis, alternative hypothesis, operating characteristics, false detection rate.

1 Introduction

Suppose that we have a statistical model for data $X \in \mathfrak{X}$ that depends upon a parameter $\theta \in \Theta$ and we want to test the null hypothesis $\mathscr{H}_0 : \theta = \theta_0$ against the alternative $\mathscr{H}_1 : \theta \neq \theta_0$ using a test statistic $T(X) \in \mathfrak{T} \subseteq 0 \cup \mathbb{R}^+$, where large values of T are taken as providing evidence against $\mathscr{H}_0$. Having observed $t = T(x)$, the evidence against $\mathscr{H}_0$ is often summarized in the P–value $p = P(T \geq t|\mathscr{H}_0)$, with small values of p being taken as undermining $\mathscr{H}_0$ in favour of $\mathscr{H}_1$. Let $H = 0$ signify that $\mathscr{H}_0$ is true and $H = 1$ signify that $\mathscr{H}_0$ is false and $\mathscr{H}_1$ is true. The implied test is equivalent to the decision rule $H(T(X))$ where

$$H(t) = \begin{cases} 0, & \text{if } p > \alpha; \\ 1, & \text{if } p \leq \alpha. \end{cases}$$

Since P–values satisfy the inequality

$$\Pr(P \leq \alpha|\mathscr{H}_0) \leq \alpha$$

for all $\alpha \in [0, 1]$, the probability of a false rejection using this decision rule is no more than α and the significance level of the test, the Type I error, is α.

Having controlled the Type I error, the classical approach to statistical testing is to try to determine the critical region so as to maximize power, $P(T \geq t|\mathscr{H}_1)$. Power calculations can be used to guide the choice of an appropriate test statistic or sample size, but they do not explicitly enter the testing process in practice. In a significance test a decision to accept or reject the null hypothesis is driven solely by an examination of the strength of evidence against $\mathscr{H}_0$, summarized in the value of p. See, *inter alia*, Cox and Hinkley (1974, Chapters 3-4). As noted by Cox and Hinkley, a serious limitation of such significance tests is that the data could be consistent or inconsistent with an alternative of little practical interest or one of great practical importance, whatever the P–value.

More recently, Moerkerke et al. (2006) have argued that the special status of the null hypothesis is often unwarranted or inappropriate. Working in the context of genetic marker selection and within the simple statistical framework of one sided tests of the difference between two mean values, Moerkerke et al. (2006) suggest that rejecting a difference in two means of a

specific magnitude can equally well justify ignoring a genetic marker in further studies as acceptance of no statistically significant difference. They therefore argue that evidence against the alternative can be equally meaningful, particularly for those who are concerned with the detection of practically important alternatives and who are prepared to specify an alternative value of scientific or practical interest. Consequently, they have proposed giving consideration to not only classical P-values, but also their counterparts derived under the alternative of interest. They suggest summarizing evidence from the perspective of both the null and the target alternative, and then balancing these.

In this paper we expand upon and generalize the ideas introduced in Moerkere et al (2006). We develop the concept of dual P–values and evidential tension, and show how they can be used to construct a practical decision rule, a *balanced* hypothesis test, that incorporates the notions of both size and power. The operational characteristics of *balanced* hypothesis tests are outlined and their relationship to conventional criteria for optimal tests is examined. The employment of *balanced* hypothesis tests as a conceptual, interpretive device is illustrated via model selection in linear regression, and their practical use is demonstrated by application to the detection of cancer-specific protein markers in mass spectroscopy.

2 Dual P–values

To begin, consider testing $\mathscr{H}_0 : \theta = \theta_0$ versus $\mathscr{H}_1 : \theta = \theta_1 \neq \theta_0$ using a test statistic T. We suppose that T is sufficient for θ and that the distribution function of T is given by $F_0(t)$ under $\mathscr{H}_0$ and $F_1(t)$ under $\mathscr{H}_1$, denoted $T \sim T_0$ when $H = 0$ and $T \sim T_1$ when $H = 1$ respectively. Also assume that F_0 and F_1 are continuous with common support $\mathfrak{T}$, respective densities $f_0 > 0$ and $f_1 > 0$, a.e. , and that T_0 is stochastically smaller than T_1. Now let

$$p_0 = Pr(T \geq t|H = 0) = Pr(T_0 \geq t) = 1 - F_0(t)$$

the classical P–value, and set

$$p_1 = Pr(T < t|H = 1) = Pr(T_1 < t) = F_1(t)\,.$$

Although small values of p_0 may represent strong evidence against $\mathscr{H}_0$ such values do not

necessarily represent strong evidence in favour of the alternative. The alternative P-value p_1, on the other hand, measures how likely it is to obtain a statistic as small as or smaller than the value observed when the alternative is assumed true. The alternative P–value corresponds, of course, to the classical P–value for testing $\mathscr{H}_1$ versus $\mathscr{H}_0$ and provides a summary of the evidence against $\mathscr{H}_1$, small values of p_1 undermining $\mathscr{H}_1$ in favour of $\mathscr{H}_0$. As suggested in Moerkere et al (2006), a large p_1 indicates an event quite likely under the alternative hypothesis and we can think of the evidence against $\mathscr{H}_0$ and in favour of $\mathscr{H}_1$ growing as p_0 becomes smaller and p_1 becomes larger. Thus, p_0 and p_1 perform a dual role and will be referred to as dual P–values.

Example 1: Let $X = \{X_1, \ldots, X_n\}$ where X_i are i.i.d. $N(\mu, \sigma^2)$ so that $\mathfrak{X} = \mathbb{R}^n$ and $\theta = (\mu, \sigma^2)$. Assume, for simplicity, that σ^2 is known and that we wish to test $\mathscr{H}_0 : \mu = \mu_0$ versus $\mathscr{H}_1 : \mu \neq \mu_0$. If λ_n is used to denote the likelihood ratio statistic, then it is easily shown that

$$T = -2\log(\lambda_n) = \frac{n}{\sigma^2}(\overline{X} - \mu_0)^2$$

where $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$. Under $\mathscr{H}_0$ the statistic $T \sim T_0 = \chi^2(1,0)$, whilst under the alternative $T \sim T_1 = \chi^2(1,\gamma)$, $\gamma = n(|\mu_1 - \mu_0|/\sigma)^2$, for any $\mu_1 \neq \mu_0$.

If we observe $T = t$ then $p_0 = Pr(\chi^2(1,0) > t)$ and $p_1 = Pr(\chi^2(1,\gamma) \leq t)$. It is well known that $Pr(\chi^2(1,\gamma) > t)$ is an increasing function of γ (Gupta and Perlman, 1974) and the traditional interpretation of this result is that the power of the test is an increasing function of n for a given $|\mu_1 - \mu_0|$, indicating the ability of a significance test based on T to discriminate more clearly between $\mathscr{H}_0$ and $\mathscr{H}_1$ as the sample size increases. From a practical perspective however, this property implies that for a given n the P–value p_1 will decrease as $|\mu_1 - \mu_0|$, the difference between the null and the alternative, increases. The latter indicates that if an alternative value of scientific or practical interest differs from the null by a considerable margin then p_1 could be small and the data need not provide evidence in favour of $\mathscr{H}_1$, even if p_0 is itself very small and the credibility of $\mathscr{H}_0$ is seriously in doubt.

This elementary example clearly demonstrates one of the basic tenants of the current approach. If the practitioner is prepared to designate a particular alternative of scientific or

practical relevance then, intuitively, we would wish to reject $\mathscr{H}_0$ in favour of $\mathscr{H}_1$ only if the observed value of T appears improbable under $\mathscr{H}_0$ whilst, at the same time, appearing quite likely under $\mathscr{H}_1$.

The following example (i) illustrates the interaction between p_0 and p_1, (ii) demonstrates the perhaps rather trite point that F_0 and F_1 only have to be known approximately, and, more importantly, (iii) it shows that our ideas and results have very broad applicability.

Example 2: Suppose that we have a statistical model for data $X = \{X_1, \ldots, X_n\}$ that depends upon a parameter $\theta = (\theta_1, \ldots, \theta_p)$ and we want to test a null hypothesis characterized by a set of $r \leq p$ linearly independent restrictions, that is, we wish to test $\mathscr{H}_0 : h(\theta) = 0$ against the alternative $\mathscr{H}_1 : h(\theta) \neq 0$.

Assume that we have an unrestricted estimator $\widehat{\theta}_n$ of theta, $\widehat{\theta}_n : \mathfrak{X} \mapsto \Theta$, such that $\text{plim}_{n\to\infty}(\widehat{\theta}_n - \theta_n^*) = 0$ where the sequence $\{\theta_n^*\}$ is interior to Θ uniformly in n. Suppose also that there exists a sequence of matrices $\Sigma(\theta_n^*)$ that are $O(1)$ and uniformly positive definite such that

$$\Sigma\left(\theta_n^*\right)^{-1/2} \sqrt{n}\left(\widehat{\theta}_n - \theta_n^*\right) \xrightarrow{\mathscr{L}} N\left(0, I_p\right) \tag{1}$$

as $n \to \infty$. The idea behind the Wald test principle (Wald, 1943) is to examine $h(\theta)$ at $\theta = \widehat{\theta}_n$ and ascertain if $\widehat{\theta}_n$, which is a consistent estimate of the true parameter value, seems to satisfy the constraint.

Let θ_0 denote the true value of θ and suppose that $h\left(\theta_0\right) = 0$. A Taylor series applied to the constraint function gives

$$h(\widehat{\theta}_n) = h(\theta_0) + \frac{\partial h(\theta_0)}{\partial \theta}(\widehat{\theta}_n - \theta_0) + o(\|\widehat{\theta}_n - \theta_0\|)$$

and therefore

$$\sqrt{n}h\left(\widehat{\theta}_n\right) = H(\theta_0)\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) + o_p(1)$$

where $H(\theta) = \partial h(\theta)/\partial \theta$ because $\sqrt{n}\|\widehat{\theta}_n - \theta_0\| = O_p(1)$ by (1) applied with $\theta_n^* = \theta_0 \ \forall \ n$. Thus, by an application of Cramér's theorem and the delta method we have

$$\left[H(\theta_0)\Sigma(\theta_0)H'(\theta_0)\right]^{-1/2} \sqrt{n}h\left(\widehat{\theta}\right) \xrightarrow{\mathscr{L}} N\left(0, I_r\right)$$

under $\mathscr{H}_0$. Let

$$T = nh(\theta)' \left[H(\theta)\Sigma(\theta)H(\theta)\right]^{-1} h(\theta)\Big|_{\theta=\widehat{\theta}_n} .$$

Slutsky's theorem and the continuos mapping theorem yield the result that under $\mathscr{H}_0$

$$T \xrightarrow{\mathscr{L}} T_0 \sim \chi^2(r,0)$$

and p_0 can be approximated by $Pr(\chi^2(r,0) > t)$.

Now consider a sequence of local alternatives $\mathscr{H}_{1n}$ given by $\theta_{1n} = \theta_0 + \zeta/\sqrt{n}$ where $\|\zeta\| < \infty$ and $h(\theta_{1n}) \neq 0$. Then

$$h(\widehat{\theta}_n) = h(\theta_{1n}) + H(\theta_{1n})(\widehat{\theta}_n - \theta_{1n}) + o(\|\widehat{\theta}_n - \theta_{1n}\|)$$

and

$$h(\theta_{1n}) = H(\theta_0)\left(\frac{\zeta}{\sqrt{n}}\right) + o(\|\frac{\zeta}{\sqrt{n}}\|).$$

Since $\|H(\theta_{1n}) - H(\theta_0)\| \to 0$ it follows from (1) with $\theta_n^* = \theta_{1n}$ that

$$\left[H(\theta_0)\Sigma(\theta_0)H'(\theta_0)\right]^{-1/2} \sqrt{n}h(\widehat{\theta}_n) = \lambda + Z_n + o_p(1)$$

where $\lambda = \left[H(\theta_0)\Sigma(\theta_0)H'(\theta_0)\right]^{-1/2} H(\theta_0)\zeta$ and $Z_n \xrightarrow{\mathscr{L}} N(0, I_r)$. From this we can conclude that

$$T = (\lambda + Z_n)'(\lambda + Z_n) + o_p(1).$$

Hence under $\mathscr{H}_{1n}$

$$T \xrightarrow{\mathscr{L}} T_1 \sim \chi^2(r,\gamma)$$

where $\gamma = \|\lambda\|^2 = \zeta' H'(\theta_0)\left[H(\theta_0)\Sigma(\theta_0)H'(\theta_0)\right]^{-1} H(\theta_0)\zeta$ and p_1 can be approximated by $Pr(\chi^2(r,\gamma) \leq t)$.

Figure 1 illustrates both P–values, p_0 and p_1, when $r = 4$ and $\gamma = 6$. If the value $T = t_1$ were to be observed then the evidence against $\mathscr{H}_0$ and in favour of $\mathscr{H}_{1n}$ might be thought of as being relatively weak because p_0 is large and p_1 is small, but if $T = t_2$ then the evidence against $\mathscr{H}_0$ and in favour of $\mathscr{H}_{1n}$ could be regarded as being quite strong because p_0 is small and p_1 correspondingly large.

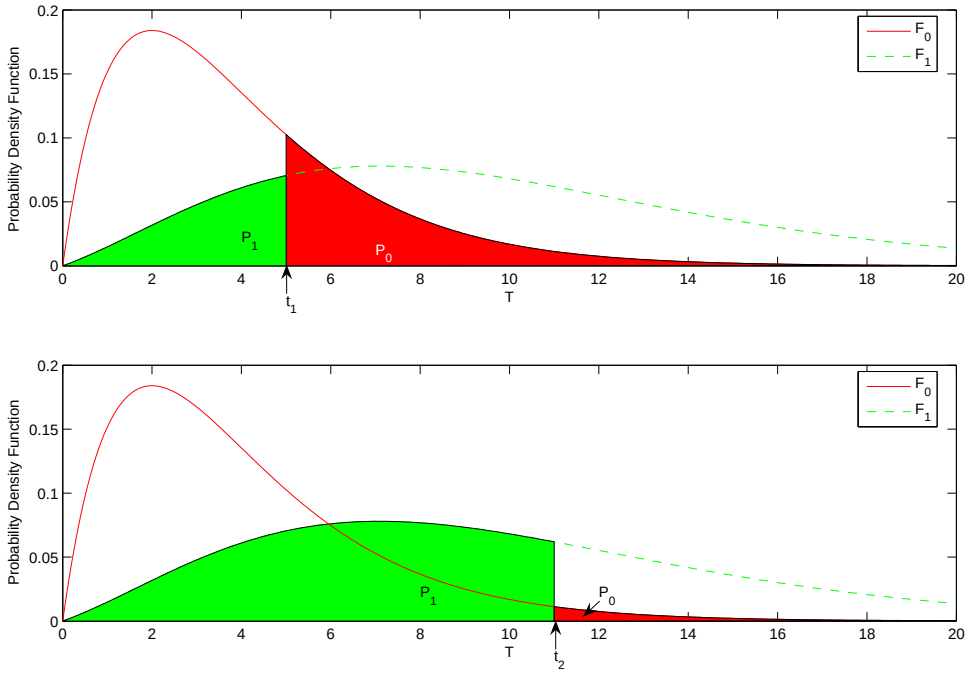


Figure 1: *Graphical presentation of dual P–values* p_0 *and* p_1.

The generic structure discussed in this example is applicable to least squares, quasi maximum likelihood estimation and M-estimators. For a discussion of different conceptual frameworks and regularity conditions that give rise to a consideration of Pitman drift (Pitman, 1979, Chapter 7) and the type of scenario discussed here see, amongst others, White (1994).

The previous examples clearly indicate that it might be useful to modify the traditional significance test by considering some kind of trade–off between the dual P–values, a trade–off that amounts to using the conventional notions of both size and power directly in the data analysis.

3 P–values and Tension

Generalizing slightly, let π_0 be the *a priori* probability that the hypothesis $\mathscr{H}_0$ is true: that is, we assume that H is a Bernoulli random variable with $Pr(H = 0) = \pi_0$ and $Pr(H = 1) = 1 - \pi_0 = \pi_1$, $0 < \pi_0 < 1$. Then $T|H \sim (1 - H) \cdot T_0 + H \cdot T_1$ and the marginal distribution of T is given by the finite mixture $\pi_0 F_0 + \pi_1 F_1$. Adopting a decision theory framework, suppose

that the consequences of our actions are characterized by the losses given in Table 1 where l_{ij}

Table 1: *Loss Table*

		Action	
		$H=0$	$H=1$
$H=0$		l_{00}	l_{01}
$H=1$		l_{10}	l_{11}

State of the World

denotes the loss incurred from deciding $\mathscr{H}_j$ obtains when in fact $\mathscr{H}_i$ holds, $i, j = 0, 1$. The loss incurred from making an incorrect decision is assumed to exceed that from making a correct one, so the regrets $r_{01} = l_{01} - l_{00} > 0$ and $r_{10} = l_{10} - l_{11} > 0$. Consider a decision rule $H_t(T)$ of the form

$$H_t(T) = \begin{cases} 0, & \text{if } T \leq t; \\ 1, & \text{if } T > t. \end{cases}$$

We seek to determine $H_t(T)$ such that the Bayes risk

$$BR = \pi_0[l_{00}Pr(T_0 \leq t) + l_{01}Pr(T_0 > t)] + \pi_1[l_{10}Pr(T_1 \leq t) + l_{11}Pr(T_1 > t)]$$

is minimized. The solution is derived from a generalization of the Neyman-Pearson lemma and is given in the following theorem, which incorporates this well known result in an unfamiliar guise.

Theorem 1 *Suppose that $T_1 = h(T_0)$. Then $h(t)$ is a continuous, increasing and differentiable function of t and BR is minimized by setting $t = t^*$ in $H_t(T)$ where*

$$\frac{f_0(h^{-1}(t^*))|dh^{-1}(t^*)/dt|}{f_0(t^*)} = k = \frac{\pi_0 r_{01}}{(1-\pi_0)r_{10}}.$$

Theorem 1 indicates how to partition $\mathfrak{T}$ in such a way that the Bayes risk is minimized and we now wish to developed a decision procedure based on $H_{t^*}(T)$ that uses the dual P–values p_0 and p_1. To this end, let $\alpha^* = Pr(T \geq t^*|H=0) = Pr(T_0 \geq t^*)$ and $\beta^* = Pr(T < t^*|H=1) = Pr(T_1 < t^*)$, and set

$$\rho^* = \frac{1-\beta^*}{1-\alpha^*}.$$

We will refer to ρ^* as evidential tension. The nomenclature recognizes that ρ^* is an increasing function of α^* and a decreasing function of β^* and, ideally, we would like to make both error

probabilities small. Since for a given sample size decreases in α^* are matched by increases in β^*, and visa–versa, ρ^* can be viewed as measuring the relative balance between them.

To assess the optimal course of action given the data, we compare the observed evidential tension with its theoretical equivalent and

$$\begin{aligned} \text{accept } \mathscr{H}_0 \text{ and reject } \mathscr{H}_1 \text{ if } \frac{1-p_1}{1-p_0} &\geq \rho^*, \\ \text{accept } \mathscr{H}_1 \text{ and reject } \mathscr{H}_0 \text{ if } \frac{1-p_1}{1-p_0} &< \rho^*. \end{aligned} \tag{2}$$

The value of ρ^* reflects the weight of evidence necessary to induce the decision maker to swap actions between $H=0$ and $H=1$ and its empirical counterpart, $(1-p_1)/(1-p_0)$, reflects the evidential content of the data. The first inequality occurs if $p_0 > \alpha^*$ and $p_1 < \beta^*$, and implies that the weight of evidence is in favour of $\mathscr{H}_0$ rather than $\mathscr{H}_1$. The second inequality occurs if $p_0 < \alpha^*$ and $p_1 > \beta^*$, implying that there is sufficient evidence to reject the null in favour of the alternative hypothesis. Following Moerkere et al (2006), we will call this a *balanced* test as it balances p_0 against p_1 via the evidential tension, treating each on an equal footing.

The *balanced* test evaluates the data from the perspective of both $\mathscr{H}_0$ and $\mathscr{H}_1$ and the decision rule is perhaps most easily implemented and understood when couched in terms of the evidential ratio

$$\mathscr{R}^*_{01} = \frac{\rho^*(1-p_0)}{(1-p_1)}. \tag{3}$$

For a given ρ^*, the ratio $\mathscr{R}^*_{01}$ increases as p_0 decreases and p_1 increases, so the larger is $\mathscr{R}^*_{01}$ the more evidence there is in favour of $\mathscr{H}_1$ rather than $\mathscr{H}_0$. If $p_0 < \alpha^*$ and $p_1 > \beta^*$ then $\mathscr{R}^*_{01} > 1$ and the observed value of the test statistic implies that the evidence in the data in favour of $\mathscr{H}_1$ rather than $\mathscr{H}_0$ is greater than that implicit in the value of ρ^*. Thus if p_0 is sufficiently small and p_1 sufficiently large $\mathscr{R}^*_{01}$ will exceed unity and the decision rule becomes:

$$H(\mathscr{R}^*_{01}) = \begin{cases} 0, & \text{if } \mathscr{R}^*_{01} \leq 1; \\ 1, & \text{if } \mathscr{R}^*_{01} > 1. \end{cases}$$

Example 3: Let $X = \{Z_1, \ldots, Z_n\} \times \{Y_1, \ldots, Y_m\}$ where Z_i are i.i.d. $N(\mu_1, \sigma^2)$ and are independent of Y_i i.i.d. $N(\mu_2, \sigma^2)$, so that $\mathfrak{X} = \mathbb{R}^n \times \mathbb{R}^m$ and $\theta = (\mu_1, \mu_2, \sigma^2)$. Assume that we wish to test $\mathscr{H}_0 : \mu_1 = \mu_2$ versus $\mathscr{H}_1 : \mu_1 \neq \mu_2$ with σ^2 being unspecified. Then the likelihood ratio

statistic is given by

$$\lambda_{n,m} = \left[\frac{\nu}{\nu + D^2}\right]^{(\nu+2)/2}$$

where $\nu = n + m - 2$ and $D = (\overline{Z} - \overline{Y})/\overline{\sigma}$,

$$\overline{\sigma}^2 = \left(\frac{1}{n} + \frac{1}{m}\right)\left(\frac{\sum_{i=1}^{n}(Z_i - \overline{Z})^2 + \sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n + m - 2}\right).$$

The likelihood ratio $\lambda_{n,m}$ is a monotonically decreasing function of $T = D^2$ and the likelihood ratio significance test is obviously equivalent to a significance test based on T.

The statistic $T \sim T_0 = F(1, \nu, 0)$ under $\mathscr{H}_0$ and the classical significance test amounts to the F–test corresponding to a two–sided t–test. Under $\mathscr{H}_1$, $T \sim T_1 = F(1, \nu, \gamma)$, $\gamma = \nu(|\mu_1 - \mu_2|/\sigma)^2$. Let $\Delta = (\mu_1 - \mu_2)/\overline{\sigma}$, the contrast measured in units of the standard error, and set $T' = (D - \Delta)^2$. Then $T' \sim T_0$ under $\mathscr{H}_1$ and $T = (\sqrt{T'} + \Delta)^2$. Applying Theorem 1 with $T_1 = h(T_0) = (\sqrt{T_0} + \Delta)^2$ and

$$f_0(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\Gamma(1/2)} \frac{(\nu t)^{-1/2}}{(\nu + t)^{(\nu+1)/2}}$$

leads to the solution of

$$\frac{(\nu + t^*)}{(\nu + (\sqrt{t^*} - \Delta)^2)} = k^{2/(\nu+1)} = c$$

for t^*. When $c = k = 1$ this gives $t^* = \Delta^2/4$, otherwise

$$t^* = \frac{\left(c|\Delta| - \sqrt{c\Delta^2 - (1-c)^2\nu}\right)^2}{(1-c)^2}.$$

Given an assigned value for Δ, based on the context plus knowledge and experience of the problem at hand, it is now a straightforward matter to calculate t^* and then ρ^* and implement a *balanced* F–test.

4 Operating Characteristics

In his discussion of frequentist inference and hypothesis testing, Welsh (1996, Section 3.4) argues that the approach to hypothesis testing that now dominates much of the literature is a hybrid of the Fisherian and Neyman-Pearson paradigms, made possible by use of the

traditional P–value. From the above discussion it is clear that the *balanced* test can be viewed similarly, a hybridization that uses the dual P–values. Given that our development has been based upon the decision theoretic considerations associated with minimization of the Bayes risk, it is perhaps natural at this point to turn to an investigation of the frequentist properties of $H(\mathscr{R}^*_{01})$, following the rationale of Rubin (1984).

Consider the behaviour of $H(\mathscr{R}^*_{01})$ as a function of t, the observed value of the statistic T, and t^*, the analytically derived value that determines the evidential tension ρ^*. If $t \leq t^*$ then $p_0 \geq \alpha^*$ and $p_1 \leq \beta^*$, implying that

$$\frac{1-p_o}{1-\alpha^*} \leq 1 \text{ and } \frac{1-\beta^*}{1-p_1} \leq 1$$

and hence that $\mathscr{R}^*_{01} \leq 1$. If, on the other hand, $t > t^*$ then $p_0 < \alpha^*$ and $p_1 > \beta^*$, which implies that

$$\frac{1-p_o}{1-\alpha^*} > 1 \text{ and } \frac{1-\beta^*}{1-p_1} > 1$$

and hence that $\mathscr{R}^*_{01} > 1$. It follows directly that

$$\begin{aligned} Pr(\mathscr{R}^*_{01} > 1|\mathscr{H}_0) &= Pr(t > t^*|H = 0) \\ &= \alpha^*. \end{aligned}$$

Lemma 1 *The* balanced *test has size α^* and is equivalent to a likelihood ratio test of $\mathscr{H}_0$ versus $\mathscr{H}_1$ based upon T and conducted at significance level α^*.*

The critical value of the likelihood ratio test of Lemma 1, k, is determined by the prior odds ratio $\pi_0/(1-\pi_0)$ and the ratio of regrets $(l_{01}-l_{00})/(l_{10}-l_{11})$ and these ultimately govern the value of α^* through k.

Lemma 1 suggests that the *balanced* test will inherit the desirable asymptotic (large sample) properties of the likelihood ratio test, as discussed in Cox and Hinkley (1974, Section 9.3) for example.

Example 2 (Revisited): If $\mathscr{H}_1$ holds then $h(\boldsymbol{\theta}_0) \neq \mathbf{0}$ and

$$\text{plim}\left\{h\left(\widehat{\theta}_n\right)'\left[H(\widehat{\theta}_n)\Sigma(\widehat{\theta}_n)H(\widehat{\theta}_n)\right]^{-1}h\left(\widehat{\theta}_n\right)\right\} > 0\,.$$

It follows that under a fixed alternative the test statistic T diverges to ∞ as $n \to \infty$. This implies that with a fixed alternative $p_1 \to 1$ as $n \to \infty$ and $\mathscr{R}^*_{01} \to \infty$. The test based on $H(\mathscr{R}^*_{01})$ will be consistent.

Turning to the finite sample properties of the *balanced* test, note first that an immediate consequence of Lemma 1 is that the probability of rejecting $\mathscr{H}_0$ when it is true will not exceed the power, that is, the test is unbiased. In particular we have

$$\begin{aligned}
\alpha^* &= Pr(\mathscr{R}^*_{01} > 1|\mathscr{H}_0) \\
&\leq Pr(\mathscr{R}^*_{01} > 1|\mathscr{H}_1) \\
&= 1 - Pr(\mathscr{R}^*_{01} \leq 1|\mathscr{H}_1) \\
&= 1 - Pr(t \leq t^*|H = 1) \\
&= 1 - \beta^*\,.
\end{aligned}$$

Since the decision rules $H(\mathscr{R}^*_{01})$ and $H_{t^*}(T)$ are equivalent it also follows from Theorem 1 that the *balanced* test is admissible (See Cox and Hinkley, 1974, pp. 431-433) . Admissibility is a relatively weak property, however, and it does not appear to be possible to say anything more general about the power of the *balanced* test analytically than is given in the following result.

Corollary 1 *The* balanced *test is a point optimal test, that is, it is equivalent to the most powerful test of size α^* for testing the simple null $\mathscr{H}_0$ against the simple alternative $\mathscr{H}_1$.*

5 Applications

5.1 Regression Modeling

Suppose that we are interested in modeling a real valued stochastic process y_t, $t \in \mathbb{N}$ using a linear regression model where the regressors are chosen from the collection $\mathscr{R} = \{x_{tk} : k =$

$1,\ldots,N\}$, $N \in \mathbb{N}$, of real-valued processes defined on the same probability space as y_t. Here $\mathbb{N}$ denotes the natural numbers (positive integers) and the variables in $\mathscr{R}$ are those deemed to be appropriate for the analysis at hand. The problem of interest is to find the subset of regressors that may be regarded as most important. In this case the set of relevant specifications contains 2^N different models where under model $\mathscr{M}_\rho$ the regressors $\{x_{tr(k)} : k = 1,\ldots,k_r\}$ enter the regression equation, that is, $\rho = \{r(1),\ldots,r(k_r)\} \subseteq \{1,\ldots,N\}$ denotes the subset of regressors in $\mathscr{R}$ that appear in the r'th model $r = 1,\ldots,2^N$.

A common approach to this problem is to consider the use of model selection criteria and one of the earliest of these is the information criterion proposed by Akaike (1974). Writing Akaike's criterion as

$$-2\max(\text{log likelihood}) + 2(\#\text{ estimated parameters}), \tag{4}$$

we find that under Gaussian assumptions the criterion function for model $\mathscr{M}_\rho$ becomes

$$AIC(\mathscr{M}_\rho) = n\log\widehat{\sigma}^2_\rho + 2(k_r+1)$$

where $n\widehat{\sigma}^2_\rho$ is the residual sum of squares achieved with that model, n being the sample size. Let $\mathbf{y} = (y_1,\ldots,y_n)'$ denote the $n\times 1$ vector of observations on the dependent variable, or regressand, and similarly define $\mathbf{X}_\rho$ to be the $n\times k_r$ observation matrix for the regressor set for model $\mathscr{M}_\rho$ with rows $\mathbf{x}'_{rt} = (x_{tr(1)},\ldots,x_{tr(k_r)})$ for $t = 1,\ldots,n$. In our case we can assume, without loss of generality, that the regressors in $\mathscr{R}$ contain no redundancies, so that $\mathbf{X}_\rho$ has full column rank. Let $\mathbf{P}_\rho = \mathbf{X}_\rho(\mathbf{X}'_\rho\mathbf{X}_\rho)^{-1}\mathbf{X}'_\rho$ denote the (prediction) operator that projects on to the space spanned by the columns of $\mathbf{X}_\rho$ and $\mathbf{R}_\rho = \mathbf{I}_n - \mathbf{P}_\rho$ the associated (residual) operator which projects on to the orthogonal complement of that space. Then $n\widehat{\sigma}^2_\rho = \|\mathbf{R}_\rho\mathbf{y}\|^2$, where for $\mathbf{x}\in\mathbb{R}^n$ $\|\mathbf{x}\|^2 = \mathbf{x}'\mathbf{x}$.

The criterion is implemented by choosing the model $\mathscr{M}_\rho = \widehat{\mathscr{M}_{AIC}}$ such that for all $r = 1,\ldots,2^N$, $AIC(\widehat{\mathscr{M}_{AIC}}) \leq AIC(\mathscr{M}_\rho)$, and since the introduction of model selection criteria of this type an extensive literature has been built up concerning their empirical and theoretical behaviour, see, *inter alia*, McQuarrie and Tsai (1998). From a heuristic viewpoint, it is apparent from 4 that $\widehat{\mathscr{M}_{AIC}}$ coincides with the model that is not rejected when tested against all other candidate models – using a test based on the likelihood ratio statistic with a critical value

determined by the value of the penalty term. In the light of Lemma 1 it seems reasonable to think that AIC might admit an interpretation via *balanced* hypothesis testing.

Consider adding the regressor x_{ti}, where $i \notin \{r(1),\ldots,r(k_r)\}$, to the model $\mathscr{M}_\rho$. Set $\mathbf{w} = \mathbf{R}_\rho(x_{1i},\ldots,x_{ni})'$ and let $\mathscr{M}_{\{\rho\cup i\}}$ denote the model containing the regressors $\{x_{tr(k)} : k = 1,\ldots,k_r\}$ and x_{ti}. Using the Frisch-Waugh-Lovell theorem it is straightforward to show that the mean squared residual of $\mathscr{M}_{\{\rho\cup i\}}$ is given by

$$\begin{aligned}\widehat{\sigma}^2_{\{\rho\cup i\}} &= \widehat{\sigma}^2_\rho - (\mathbf{w}'\mathbf{y})^2/n(\mathbf{w}'\mathbf{w}) \\ &= \widehat{\sigma}^2_\rho \left[1 + \frac{\tau^2_\rho(\mathbf{w})}{(n-k_r-1)}\right]^{-1} \end{aligned} \tag{5}$$

where

$$\tau^2_\rho(\mathbf{w}) = \frac{(n-k_r-1)(\mathbf{w}'\mathbf{y})^2}{n\widehat{\sigma}^2_{\{\rho\cup i\}}(\mathbf{w}'\mathbf{w})} = \left[\frac{(\mathbf{w}'\mathbf{y})}{(\mathbf{w}'\mathbf{w})}\right]^2 \left[\frac{n\widehat{\sigma}^2_{\{\rho\cup i\}}}{(n-k_r-1)(\mathbf{w}'\mathbf{w})}\right]^{-1},$$

the square of the t-statistic for testing the significance of the regressor x_{ti}. Substituting expression 5 into $AIC(\mathscr{M}_{\{\rho\cup i\}})$ it is readily deduced that $AIC(\mathscr{M}_{\{\rho\cup i\}}) \leq AIC(\mathscr{M}_\rho)$ if and only if

$$-n\log[1 + \frac{\tau^2_\rho(\mathbf{w})}{(n-k_r-1)}] + 2 \leq 0\,.$$

Thus $\mathscr{M}_{\{\rho\cup i\}}$ will be preferred to $\mathscr{M}_\rho$ whenever

$$\tau^2_\rho(\mathbf{w}) \geq (n-k_r-1)[\exp(2/n)-1]\,.$$

Now let β denote the regression coefficient in the regression of y_t on w_t and set

$$t^*_{AIC} = (n-k_r-1)[\exp(2/n)-1]\,.$$

Under the null hypothesis that $\mathbf{y}$ is orthogonal to $\mathbf{w}$, i.e. $\mathscr{H}_0 : \beta = 0$, $T = \tau^2_\rho(\mathbf{w}) \sim T_0 = F(1,(n-k_r-1),0)$. From a direct adaptation of the arguments employed in *Example 3* it follows that

$$\begin{aligned}\alpha^*_{AIC} &= Pr(T_0 \geq t^*_{AIC}) \\ &= Pr(F(1,(n-k_r-1),0) \geq (n-k_r-1)[\exp(2/n)-1])\end{aligned}$$

implicitly defines the significance level of a *balanced* F-test of $\mathscr{H}_0 : \beta = 0$ against the alternative $\mathscr{H}_1 : \beta = \Delta_{AIC}\overline{\sigma}$ where $\Delta_{AIC} = 2\sqrt{t^*_{AIC}}$ and

$$\overline{\sigma}^2 = n\widehat{\sigma}^2_{\{\rho \cup i\}}/(n - k_r - 1)(\mathbf{w}'\mathbf{w}).$$

This establishes the relationship between AIC and the concepts used in the construction of a *balanced* hypothesis test. It indicates that AIC corresponds to a point optimal test of $\mathscr{H}_0 : \beta = 0$ where the implicit value of β under the alternative and the implied size of the test are functions of the standard error, n and k_r (Corollary 1).

Figures 2 and 3 present graphs of α^*_{AIC} and Δ_{AIC} for different values of n and k_r. The figures also plot corresponding values for BIC (Schwarz, 1978) and HQ (Hannan and Quinn, 1979) where, via developments that parallel those employed above, $\alpha^*_{BIC} = Pr(F(1,(n-k_r-1),0) \geq t^*_{BIC})$ and $\Delta_{BIC} = 2\sqrt{t^*_{BIC}}$ with

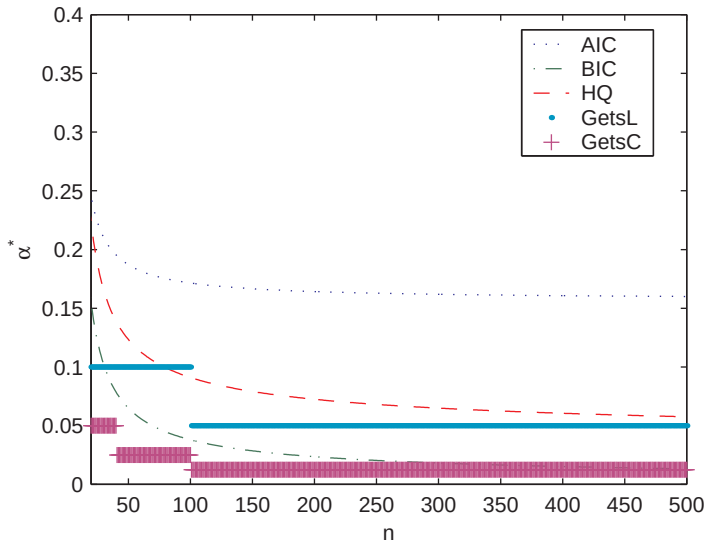
$$t^*_{BIC} = (n - k_r - 1)[n^{1/n} - 1],$$

and $\alpha^*_{HQ} = Pr(F(1,(n-k_r-1),0) \geq t^*_{HQ})$ and $\Delta_{HQ} = 2\sqrt{t^*_{HQ}}$ with

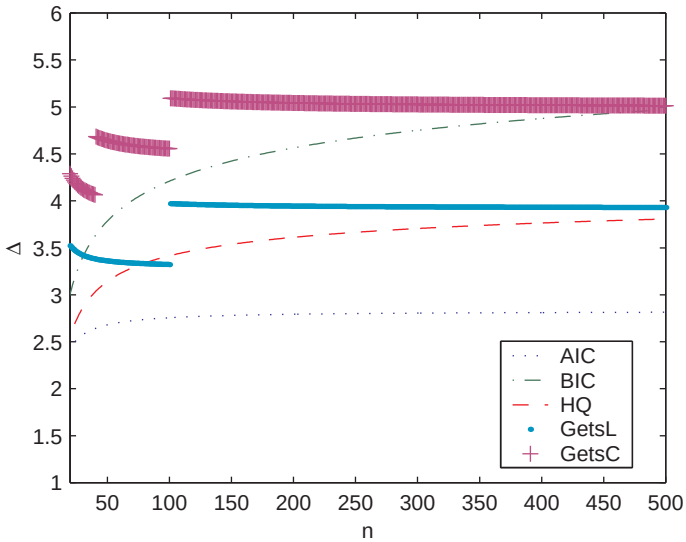
$$t^*_{HQ} = (n - k_r - 1)[(\log n)^{2/n} - 1].$$

Figures 2a and 3a indicate that for a broad range of values of k_r and n the implicit size of AIC falls roughly in the interval 30% to 16%. This is in accord with observations made by Sawa (1978). For BIC, however, α^* lies in the interval 16% to 1.25%, and the implied size of HQ falls about half way between that of AIC and BIC. For the sample sizes considered here the values of α^* implicit under AIC are much larger than conventional significance levels and only BIC and HQ generate values for their implied size that resemble the type of significance levels commonly used in practice. In fact, whereas both t^*_{BIC} and t^*_{HQ} diverge as $n \to \infty$, so α^*_{BIC} and α^*_{HQ} converge to zero asymptotically, $t^*_{AIC} \to 2$ as $n \to \infty$ and α^*_{AIC} approaches 0.1573. Although both $\alpha^*_{BIC} \to 0$ and $\alpha^*_{HQ} \to 0$ as $n \to \infty$, there are still notable differences in the implied sizes over the range of sample sizes illustrated.

In a discussion of the relationship of model selection criteria to the general-to-specific (*Gets*) least squares modeling strategy Campos et al. (2003) use the link between significance tests



(a) *Implicit Size:* α^*



(b) *Implicit Alternative in Standardized Units:* Δ

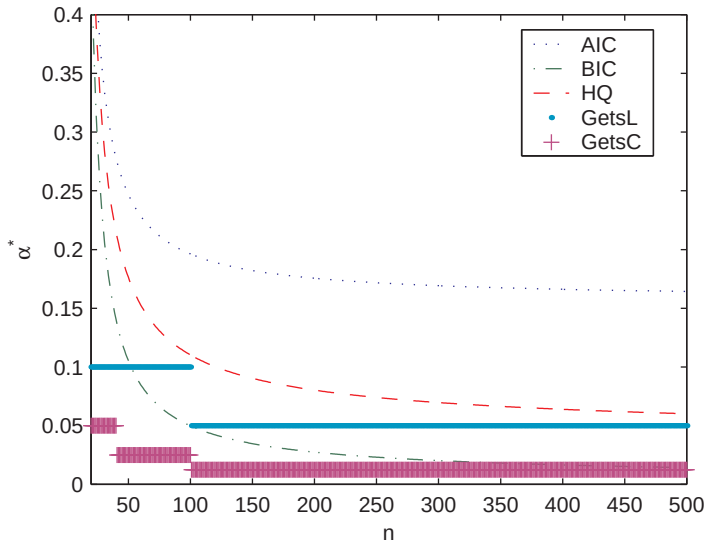
Figure 2: *Regression Modeling Balanced Hypothesis Tests,* $k_r = 5$.

and selection criteria to argue that *Gets* will be consistent. The assigned significance levels used in the *Gets* liberal ($GetsL$) and conservative ($GetsC$) strategies are also plotted in Figures 2a and 3a. These are based on conventional significance levels and decrease in discrete steps at designated sample sizes. If we denote the $GetsL$ and $GetsC$ assigned significance levels by α_L^* and α_C^*, respectively, then the size profiles indicate that whereas $GetsL$ tends to lie between BIC and HQ with $\alpha_{BIC}^* \leq \alpha_L^* \leq \alpha_{HQ}^*$ for n sufficiently large, $GetsC$ falls below all three information criteria for all n. Both *Gets* values are noticeably smaller than those of the information criteria when n is small, suggesting that they will behave very differently from the information criteria in small samples. As n increases however, $GetsL$ and HQ converge, and $\alpha_{BIC}^* - \alpha_C^* \to 0$.

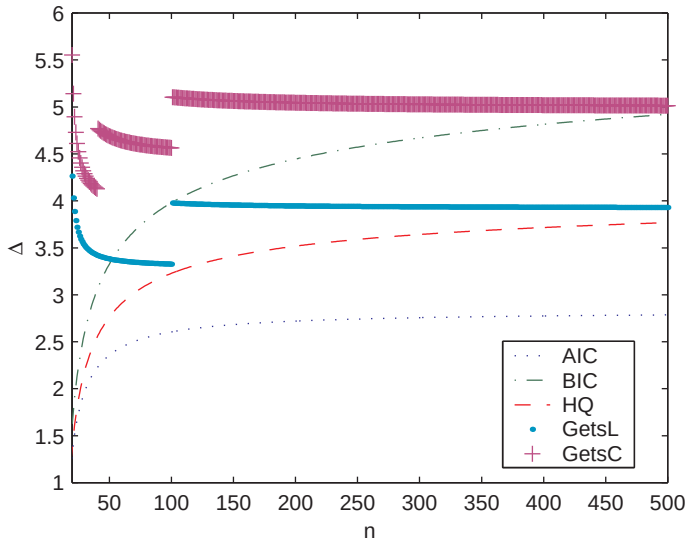
Each criterion seeks to balance the chances of omitting regressors that matter against including variables which are irrelevant and the implied alternative provides a simple characterization of how this is done. For $GetsL$ and $GetsC$, given α_L^* and α_C^*, we can calculate the implied values of t_L^* and t_C^* as the $(1-\alpha_L^*)100\%$ and $(1-\alpha_C^*)100\%$ percentile points of the $F(1,(n-k_r-1),0)$ distribution. The implied alternatives in units of the standard error are then $\Delta_L = 2\sqrt{t_L^*}$ and $\Delta_C = 2\sqrt{t_C^*}$. The values of Δ_L and Δ_C are plotted in Figures 2b and 3b. From Figures 2b and 3b we can see that AIC is balancing the null of zero against an implied alternative that never exceeds $2\sqrt{2}$ standard errors. For HQ and $GetsL$, however, Δ converges to about $2\sqrt{4}$ when $n = 500$ and for BIC and $GetsC$ Δ converges to about $2\sqrt{6}$ at this sample size. Each procedure can be interpreted as conducting a *balanced* hypothesis test and therefore each one corresponds to a point optimal test, but they are implicitly balancing the null against different values of the alternative. The larger the value under the implied alternative the more stringent the procedure and the more parsimonious the model chosen by that procedure is likely to be.

5.2 Protein Markers in Mass Spectroscopy

During the last few years there has been much interest in the use of mass spectroscopy as a tool for discrimination and screening of cancer patients. The data is collected using a surface-enhanced laser desorption/ionization system (Thiele, 2003; Banks, 2003) and consists of the proteomic spectrum of individual patients. Each spectrum gives the relative amplitude measured on a grid of mass/charge (μz) values that represent protein markers, only some of which



(a) *Implicit Size:* α^*



(b) *Implicit Alternative in Standardized Units:* Δ

Figure 3: *Regression Modeling Balanced Hypothesis Tests,* $k_r = 15$.

Test procedure	Ovarian cancer	Prostate cancer
F- test	4907	27117
bF–test	3451	22741
Total # μz values	15154	48538

Table 2: *Number of significant protein markers*

are thought to be important in diagnosing the decease. The preliminary stage of any analysis therefore consists of the identification of significant markers, see Petricoin et al. (2002) and Zhu et al. (2003) for examples of the analysis of ovarian cancer patients, and Adam et al. (2003) for prostate cancer patients.

Here we consider both ovarian cancer and prostate cancer data sets. The ovarian cancer data set contains the spectra of 100 ovarian cancer patients and 116 healthy controls (including 16 individuals with benign tumors), with each spectrum evaluated at 15154 μz values. The prostate cancer data set contains the spectra of 324 individuals, 167 of whom have prostate cancer and 157 of whom are healthy individuals (including 77 individuals with benign tumors). Here each spectrum is evaluated at 48538 μz values. (The ovarian cancer data set was downloaded from `clinicalproteomics.steem.com` and the prostate cancer data set from `www.evms.edu/vpc/seldi/`.)

The raw data sets have far more μz values (variables) than individuals (observations), but many of the μz values may represent protein markers that are uninformative. We therefore wish to consider testing the significance of the μz values or markers. Let $\mu_1 - \mu_2$ be the mean difference between the relative amplitudes of the spectra of the healthy individuals and the spectra of the cancer patients for a specific marker. Then the null and alternative hypotheses of interest are $\mathscr{H}_0 : \mu_1 = \mu_2$ versus $\mathscr{H}_1 : \mu_1 \neq \mu_2$ and we can consider testing $\mathscr{H}_0$ against $\mathscr{H}_1$ using a classical F–test and a *balanced* F–test (bF–test), as previously described in Example 3.

For each data set we have two samples, $n = 100$ cancer patients and $m = 116$ normal individuals for the ovarian cancer data, and $n = 167$ cancer patients and $m = 157$ normal individuals for the prostate cancer data, and we have carried out an F–test and a bF–test at each μz value. For the F–test the level of significance was fixed at $\alpha = 0.01$. The *balanced* test was calculated by assigning the value $\Delta = 6$ to the contrast parameter and assuming that the null and alternative hypotheses were equally likely *a priori* and $r_{10} = r_{01}$, so that $k = 1$. Table

(2) shows the number of μz values for which the null hypothesis is rejected. The *balanced* test has clearly nominated considerably fewer μz values as representing significant protein markers than the traditional F–test.

Finally, Table (2) has been constructed by conducting a large number of identical tests, $h = 15154$ for the ovarian cancer data and $h = 48538$ for the prostate cancer data. When many tests are being performed it is commonly suggested that the overall probability of making a false discovery should be controlled to account for the multiple testing. Controlling the family–wise error rate is known to lead to procedures that lack power, however, an unfortunate feature since, as here, we envisage using the test in an exploratory analysis in which interest is focused on finding significant results among many applications of the test. Following the pioneering work of Benjamini and Hochberg (1995) a number of authors have therefore suggested that, in such circumstances, an important operating characteristic of the test to examine is the rate of false positives.

Suppose that h identical hypothesis tests are performed with the statistics $T_1, ..., T_h$. Let R denote the total number of hypotheses that are rejected and let V be the number of true null hypotheses that are erroneously rejected. Storey (2003) defines the positive false discovery rate (pFDR) to be the expected proportion of false discoveries, conditional on at least one positive finding having occurred,

$$pFDR = E\left[\frac{V}{R}\,|R>0\right].$$

Assume that $(H_i, T_i,)$ are i.i.d. random variables where H_i is Bernoulli$(1-\pi_0)$ and $T_i|H_i \sim (1-H_i)\cdot T_0 + H_i \cdot T_1$, for $i = 1,\ldots,h$. Then (Storey, 2003, Theorem 1)

$$pFDR = \frac{\pi_0 Pr(T_0 > t')}{\pi_0 Pr(T_0 > t') + (1-\pi_0)Pr(T_1 > t')} = \frac{\pi_0\alpha'}{\pi_0\alpha' + (1-\pi_0)(1-\beta')}$$

where t' is the critical value used in the application of the tests, and α' and β' are the corresponding error rates. Evaluating $pFDR$ for the two tests using the parameter values previously assigned leads to the conclusion that the conventional F–test is at least 3.6 times more likely to indicate a false positive than is the *balanced* version.

Appendix: Proofs

Proof Theorem 1: By assumption $F_0(t) = F_1(h(t))$ and T_1 is stochastically larger than T_0, so $h(t) \geq t$. If $t > t'$ then

$$F_0(t) > F_0(t') \equiv F_1(h(t)) > F_1(h(t')) \Rightarrow h(t) > h(t').$$

Moreover,

$$F_0(t) - F_0(t') \equiv F_1(h(t)) - F_1(h(t')) = f_1(\bar{h})(h(t) - h(t'))$$

where $\bar{h} = h(t') + \tau(h(t) - h(t'))$, $0 \leq \tau \leq 1$, implying that $h(t) - h(t') \to 0$ as $t \to t'$ since $f_1(\bar{h}) > 0$ and $F_0(t)$ is continuous. It is now straightforward to verify that

$$\lim_{t \to t'} \frac{h(t) - h(t')}{t - t'} = \frac{f_0(t)}{f_1(h(t))},$$

completing the proof of the first part of the theorem.

Applying a direct generalization of the Neyman-Pearson lemma, Rao (1965, Section 7a.2, Lemma 4) it follows that BR is minimized by the decision rule that sets $H = 0$ if $f_1(t) \leq k f_0(t)$ and $H = 1$ otherwise. Let $B_k = \{t : f_1(t) > k f_0(t)\}$. Assume that $k \geq 1$ and suppose that $[0, t^*) \subseteq B_k$ where $f_1(t^*) = k f_0(t^*)$. Then $F_1(t) > k F_0(t) \geq F_0(t)$ and $1 - F_1(t) \leq 1 - F_0(t)$ for all $t \in [0, t^*)$, contradicting the assumption that T_1 is stochastically larger than T_0. Similarly, assuming $k < 1$ and $[t^*, \infty) \nsubseteq B_k$ implies that $1 - F_1(t) \leq k(1 - F_0(t)) < 1 - F_0(t)$ for all $t \in [t^*, \infty)$, leading to the same contradiction. Thus we can conclude that $[t^*, \infty) = B_k$ and the final statement of the theorem follows by noting that

$$f_1(t) = \frac{f_0(h^{-1}(t))}{|dh(t)/dt|} = f_0(h^{-1}(t))|dh^{-1}(t)/dt|. \quad \blacksquare$$

Proof Lemma 1: From the discussion preceding the lemma we can see that $H(\mathscr{R}^*_{01})$ is equivalent to $H_{t^*}(T)$ where $f_1(t^*) = k f_0(t^*)$, $k = \{\pi_0/(1-\pi_0)\}\{(l_{01} - l_{00})/(l_{10} - l_{11})\}$, and, as we have just shown, $[t^*, \infty) = B_k = \{t : f_1(t) > k f_0(t)\}$. Hence $H(\mathscr{R}^*_{01})$ is equivalent to the likelihood ratio test based on T with critical region B_k and $Pr(B_k|\mathscr{H}_0) = Pr(t > t^*|H = 0) = \alpha^*$. Conversely, for any $k > 0$ the set B_k defines a likelihood ratio critical region of size α^*. Moreover, we can always find a prior probability π_0 and regrets $r_{01} = l_{01} - l_{00}$ and $r_{10} = l_{10} - l_{11}$ such that $\pi_0 r_{01} = k(1 - \pi_0) r_{10}$ and the likelihood ratio test is equivalent to $H_{t^*}(T)$ and the

corresponding $H(\mathscr{R}^*_{01})$. ∎

References

Adam, B. L., Qu, Y., Davis, J. W., d. Ward, M., Clements, M. A., Cazares, L. H., Sammes, O. J., Schellhammer, P. F., Yasui, Y., Feng, Z., and Wright, J. G. L. W. (2003), "Serum Protein fingerprinting coupled with a pattern matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men," *Cancer Research*, 63, 3609–3614.

Akaike, H. . (1974), "A new look at the statistical model identification," *IEEE–Transactions on Automatic Control*, AC–19, 716–723.

Banks, D. (2003), "Proteomics: A Frontier between Genomics and Metabolomics," *Chance*, 16, 6–7.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.

Campos, J., Hendry, D. F., and Krolzig, H. M. . (2003), "Consistent model selection by an automatic Gets approach," *Oxford Bulletin of Economics and Statistics*, 65, 803–819.

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

Gupta, S. D., and Perlman, M. D. (1974), "Power of the Noncentral F-Test: Effect of Additional Variates on Hotelling's T^2-Test," *Journal of the American Statistical Association*, 69(345), 174–180.

Hannan, E. J., and Quinn, B. G. (1979), "The Determination of the Order of an Autoregression," *Journal of Royal Statistical Society*, B 41, 190–195.

McQuarrie, A. D. R., and Tsai, C.-L. (1998), *Regression and Time Series Model Selection* World Scientific Publishing Company: Singapore.

Moerkerke, B., Goetghebeur, E., De Riek, J., and Roldan-Ruiz, I. (2006), "Significance And Impotence: Towards a Balanced View of The Null and the Alternative Hypotheses in Marker Selection for Plant Breeding," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(1), 61–79.

Petricoin, E. F. I., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002), "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, 359, 572–577.

Pitman, E. J. G. (1979), *Some Basic Theory for Statistical Inference* Chapman and Hall.

Rao, C. R. (1965), *Linear Statistical Inference and its Applications*, New York: J. Wiley.

Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician," *The Annals of Statistics*, 12(4), 1151–1172.

Sawa, T. (1978), "Information Criteria for Discriminating Among Alternative Regression Models," *Econometrica*, 46, 1273–1291.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Storey, J. D. (2003), "The positive false discovery rate: a Bayesian interpretation and the q-value," *The Annals of Statistics*, 31(6), 2013–2035.

Thiele, H. (2003), "Proteomics: Mass Spectrometry and Bioinformatics in Proteomics," *Chance*, 16, 29–36.

Wald, A. (1943), "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," *Transactions of the American Mathematical Society*, 54(3), 426–482.

Welsh, A. H. (1996), *Aspects of Statistical Inference* John Wiley, New York.

White, H. (1994), *Estimation, Inference and Specification Analysis* Cambridge University Press.

Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J., and Kovash, J. S. (2003), "Detection of cancer-specific markers amid massive mass spectral data," *PNAS*, 100, 14666–14671.