ELSEVIER

Journal of Statistical Planning and Inference 108 (2002) 3–21

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Hypothesis testing: a model selection approach

M. Cubedo, J.M. Oller*,[1]

Departament d'Estadística, Universitat de Barcelona, Avgda. Diagonal 645, 08028 Barcelona, Spain

Abstract

In many cases when scientists use hypothesis tests they are fully aware that the statistical model itself, and a fortiori the null hypothesis, is only a partial description of reality, and, from a logical point of view, it is not strictly true. In particular, this applies to *black-box* models—those most frequently used in statistics. From this point of view it does not seem very reasonable to choose between hypotheses in terms of their probability of error or, even worse, their conditional probabilities of error, since these indexes are too rough to be useful.

In this paper a wide range of classical hypothesis testing problems are examined from the point of view of model selection: the acceptance or rejection of the null hypothesis of such a problem is considered in terms of the selection of the most appropriate model, when one is nested inside the other. There are several reasons why a (strictly speaking false) model might be developed: it allows us to synthesize data variability, while maintaining a reasonable fit with observed data, and also it facilitates prediction of new phenomena with a certain degree of accuracy. Quite simply, modeling makes reality more readily understandable.

In this context, some problems of the classical hypothesis testing approach are discussed and several alternatives are considered, in developing methods, based on statistical estimation theory, for nested-model selection and a wide class of hypothesis testing problems. These methods are applied to various situations and compared with classical techniques. Finally, a number of questions concerning the nature of statistical inference are posed.

MSC: 62F10; 62B10; 62A99

Keywords: Hypothesis testing; Model selection; Rao distance; Information metric

1. Introduction

Any scientific discourse is concerned with a part of reality and aims to provide statements which agree with that, but all the statements are declared in a reference

* Corresponding author.
E-mail address: oller@bio.ub.es (J.M. Oller).
[1] This work is supported by DGICYT grant (Spain), PB96-1004-C02-01 and 1999SGR00059.

0378-3758/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0378-3758(02)00267-7

language, which intrinsically limits the ability to fit this reality. In spite of this limitation, it is nevertheless convenient to carry out this discussion in this general context *as if* the statements were truthful, otherwise no discourse would be possible. Consider, for instance, the conceptual representation of any direction, in physical space, through a real line. It is clear that the richer the language, the more able it is to fit the reality, but, frequently in the study of a given part of the reality, it is useful to introduce some conceptual simplification in order to represent the basic features of the problem in a more readily understandable way, although in so doing we move away from the reality. From this point of view, the basic question of statistical inference is not: *What is the probability that a given statement is true*? since it is known that any assertion is not strictly true, rather the basic question is: *How far does this statement lie from a true statement*? provided a distance to measure this were previously supplied. In other words, a formal structure for a certain class of inductive reasoning should be based more on a measure of *closeness to the* approximate *truth* than on *probability*. Observe that, when adopting this approach, the strict falseness of any statement of the reference language does not invalidate these considerations provided this language is rich enough to fit the reality for a given purpose, i.e., all we need assume is that the distance to the approximate truth is very similar to the distance to the real truth.

In the present paper we consider hypothesis testing problems in the context of parametric statistical models when the dimension corresponding to the null hypothesis is strictly smaller than the dimension of the whole model. We aim to solve problems of this kind by adopting methods which supply us with precise explanation as to why sometimes we choose the simplest model (null hypothesis) although it is not truthful.

Let us introduce some notation in order to establish the framework of the discussion. Let χ be a sample space, a a σ-algebra of subsets of χ and μ a positive measure on the measurable space (χ, a). A *parametric statistical model* may be defined as the triple $\{(\chi, a, \mu); \Theta; f\}$, where (χ, a, μ) is a measure space, Θ, also called the *parameter space*, is a manifold, and f, the *model function*, is a measurable map, $f : \chi \times \Theta \to \mathbb{R}$ such that $f(x, \theta) \geqslant 0$ and $P_\theta(\mathrm{d}x) = f(x, \theta)\mu(\mathrm{d}x)$ is a probability measure on (χ, a), $\forall \theta \in \Theta$. We shall refer to μ as the *reference measure* of the model.

In general, Θ can be any manifold modeled on any convenient model space, although in many applications Θ will be an n-dimensional C^∞ real manifold, Hausdorff and connected. Additionally, we shall assume that function f satisfies certain regularity conditions to guarantee the existence of the Fisher information matrix, being this matrix positive definite over all Θ. It is well known that in this case the parameter space has a natural Riemannian manifold structure, induced by its information metric. For further details, see Rao (1945), Atkinson and Mitchell (1981), Burbea and Rao (1982), Amari (1985), Burbea (1986) and also Oller and Corcuera (1995) among many others.

Let us consider the testing problem $\mathrm{H}_0 : \theta \in \Theta_0$ versus $\mathrm{H}_1 : \theta \notin \Theta_0$, where θ stands for the parameter corresponding to the (approximate) true probability measure and Θ_0 is a r-dimensional C^∞ real submanifold embedded in Θ, with $0 \leqslant \dim \Theta_0 < \dim \Theta$ and let x be a random sample $x = (x_1, \ldots, x_k) \in \chi^k$ of size $k \in \mathbb{N}$. In building a test, a naive approach might pose the problem in the following terms: given a random sample

x, what is the *probability that* H_0 *is true*?, and then select the hypothesis on the basis of this probability. But to answer this question we need to know a *prior distribution* on Θ, say Π, and to follow a standard Bayes approach.

In spite of this problem, since the parameter space has a manifold structure, in most of the cases it is plausible to assume that the prior measure is dominated by any measure on Θ induced, through a coordinate system, by the Lebesgue measure on $\mathbb{R}^n$, or, equivalently, by the Riemannian measure dV corresponding to the *information metric*, for further details see Jeffreys (1946) or the above-mentioned references. We shall use this Riemannian measure as a *reference* to standardize the mathematical expressions.

Notice that this assumption is compatible with an *almost* total absence of prior knowledge about the true probability measure, *less knowledge* in fact *than* the assumption of any non-informative prior. Under this hypothesis, clearly, we have a priori $\Pi(H_0) = 0$, and a posteriori $P(H_0|x) = 0$ again zero!, since $\dim \Theta_0 < \dim \Theta$, and thus Θ_0 is a zero Riemannian measure set. Observe that this conclusion has not been obtained by working with a *particular* prior: we allow the prior to be arbitrary among probability measures on Θ equivalent to dV.

Moreover, from a purely applied point of view, *experience shows* that with hypothesis testing problems of this kind, i.e., $\dim \Theta_0 < \dim \Theta$, once the significance level of the test is fixed, *the null hypothesis will* (almost) *always be rejected* when we work with *a large enough sample size*. Thus, from this point of view, which is the philosophical meaning in the acceptance of the null hypothesis? since we know, under mild assumptions, that it is almost surely false, *independently* of any particular sample. These considerations have been pointed out in the literature by several authors see, for instance, Berger (1985).

If we merely seek to imply that *the null hypothesis is approximately true* then we are more in need of a *measure of fit* between hypothesis and data than a measure of the *propensity* of a hypothesis to be truth (as was the case of the significance chi-square test used at the beginning of the present century). An interesting discussion concerning scientific hypotheses can be found in Good (1988). Other interesting discussions and criticisms of the interpretation of the hypothesis tests can be found in Rollal (1997) and Harlow et al. (1997).

Notice that the standard procedures, such as fixing a significance level (according to sample size or not), or computing a tail-probability, appear in a natural way when we consider the classical Neymann–Pearson problem: the null and the alternative hypotheses are *simple*. In that case, any *reasonable* prior leads to a positive posterior probability for the null hypothesis and a convenient way to minimize the probability of error, under an almost total absence of knowledge about the prior distribution, is to maximize the power of a test once a significance level has been fixed. But, when the alternative is a composite non-countable hypothesis, to proceed in the same way, *by analogy*, does not clarify the advantage of accepting the null hypothesis.

In the present paper, the class of hypothesis testing problems considered is viewed as the selection of an appropriate model, when one is nested inside the other. This is in fact a slightly different problem from a logical point of view, but from the above discussion, this approach will clarify the way to be followed.

A task which must be conducted before developing any model selection statistical method is to make explicit which are the advantages of using a statistical model as it is true. In the present paper, we focus our attention on the ability of statistical models to *facilitate the estimation* of the probabilistic mechanism by means of which a particular sample has been obtained. Therefore, we shall need to evaluate the goodness of an estimate of the underlying probability measure associated with a statistical experiment. This evaluation should be based on the magnitude of the errors corresponding to the estimated probability of different observable algebra events. A reasonable way to supply a quantitative scalar measure of these errors is through a distance between probability measures.

There are several distances that could be used, but since we are dealing with regular statistical models, as we have mentioned earlier, a natural candidate for measuring the deviations of the estimates is the *Rao distance*. This distance was introduced by Rao (1945) through the information matrix, which defines a Riemannian structure over the parameter space of the statistical model. The resulting differential metric, also called the *information metric* is given by

$$\mathrm{d}s^2 = \sum_{i,j=1}^{n} g_{ij}(\theta)\,\mathrm{d}\theta^i\,\mathrm{d}\theta^j,$$

where $g_{ij}(\theta)$ are the coefficients of the Fisher information matrix, and the corresponding Riemannian distance will be a measure of dissimilarity between two probability distributions. For further details see the aforementioned references and also Čencov (1982), and Oller and Corcuera (1995). It might also be natural to use the *Hellinger* or the *Bhattacharyya distances* which, locally, give the information metric up to a proportionality constant. In the present paper though we are going to use the Rao distance, mainly on account of its inherent beauty and also because the corresponding topology thus induced is always the same as the parameter space topology. This would not, however, be the case, for instance, of the Hellinger or the Bhattacharyya distances.

2. Alternative approaches

A key idea, though perhaps initially rather imprecise, in understanding the following discussion is that to accept the null hypothesis (or submodel) should mean that under certain conditions (small enough sample size), we can, *in some sense*, make better estimates of the true probability measure if we restrict the estimates to the submodel rather than if we use the whole model. Let us pose the problem in the following terms: according to the information obtained from a k-size sample x (*actual sample*) what then can we say to those who are going to work with the same (or equivalent) statistical population when handling information from samples of size m (*potential samples*)? The *actual* sample provides the available information, while the *potential* samples fix the *resolving power* that is going to be used in the future, that is, we are going to take a decision depending on the size m of the potential samples.

A simple example can help us: let us suppose that we are performing quality control at a dice factory, and assume that for each dice we have a sample of k results.

Moreover, we know that the dice are going to be used in a game of chance that involves 30 throws. The relevant question is not whether a die is *perfect*: it is not possible to build a perfect die. The relevant question is: which model will be more accurate in describing the game's outcome: a model of a perfect die (which is certainly not true) or a model based on the relative frequencies of the 30 die throws (potential sample, of size $m=30$). We have to decide between both questions based on a previous sample (actual sample, of size k). Clearly, a reasonable answer should be based on both sample sizes.

Partial answers to this question can be given, by developing a statistical method which *requires* the acceptance of an *estimator* as being suitable for the whole model, for instance, the maximum-likelihood estimator, although others may be selected. In other words, the proposed answers will be *dependent* on the corresponding estimator. Hereafter, let $T_m : \chi^m \to \Theta$ be the proposed estimator corresponding to a sample of size m, for the whole model.

To begin with, let us think *as* if the parameter corresponding to the true probability measure (or density) were known, say θ. Thus, a *reasonable* measure of the error of an estimate will be given by $\rho^2(T_m(x), \theta)$, while the error corresponding to the null hypothesis (or submodel) will be measured by

$$\rho^2(\Theta_0, \theta) = \inf\{\rho^2(\eta, \theta) : \eta \in \Theta_0\}.$$

We shall propose two ways of judging the convenience of one model with respect to another. The first one, which will be referred to hereafter as the $\mathfrak{p}$—*rule*, is as follows:

We either reject the null hypothesis, or we choose the whole model Θ if

$$\mathfrak{p} = P(\rho^2(T_m(x), \theta) < \rho^2(\Theta_0, \theta)) - P(\rho^2(T_m(x), \theta) > \rho^2(\Theta_0, \theta)) > 0$$

otherwise we accept the simplest model Θ_0. Observe that if $\rho^2(T_m(x), \theta)$ has an absolutely continuous distribution, this rule is equivalent to

$$\mathfrak{p} = 2P(\rho^2(T_m(x), \theta) < \rho^2(\Theta_0, \theta)) - 1 > 0 \tag{1}$$

i.e., if we measure the error in the estimation of the probabilistic mechanism that generates the sample through $\rho^2(T_m(x), \theta)$, we choose Θ_0 if it is strictly more probable that this error is bigger than the error corresponding to Θ_0, measured as $\rho^2(\Theta_0, \theta)$, than viceversa.

The other way that we shall consider here and which will be referred hereafter to as the $\mathfrak{m}$—*rule*, is as follows.

We either reject the null hypothesis, or we choose the whole model Θ if

$$\mathfrak{m} = \rho^2(\Theta_0, \theta) - E(\rho^2(T_m(x), \theta)) > 0 \tag{2}$$

otherwise we accept the simplest model Θ_0, i.e., if the error in the estimation of the probabilistic mechanism that generates the sample is measured through $\rho^2(T_m(x), \theta)$, we choose Θ_0 if the average of these errors is bigger than or equal to the error corresponding to Θ_0, measured as $\rho^2(\Theta_0, \theta)$.

The first rule is invariant under monotone transformations of the Rao distance. Therefore, it is less dependent on the underlying metric, or *loss function* used, while the second, based on moments, is sometimes easier to handle.

Notice that for any consistent estimator T_m, if $\rho^2(\Theta_0,\theta)>0$, which occurs almost surely under our general assumptions, then

$$\lim_{m\to\infty} P(\rho^2(T_m(x),\theta)<\rho^2(\Theta_0,\theta))=1$$

and for any L^2-consistent (with respect to the Riemannian metric) estimator T_m,

$$\lim_{m\to\infty} E(\rho^2(T_m(x),\theta))=0$$

always rejecting the null hypothesis for large sample sizes: in this case we shall have enough information to be able to determine that the null hypothesis is wrong.

On the other hand, if $\rho^2(\Theta_0,\theta)=0$, which is the case with probability zero under our assumptions, then $\mathfrak{p}<0$ and $\mathfrak{m}\leqslant 0$, always accepting the null hypothesis (submodel), for all $m\in\mathbb{N}$.

Observe that the *decisions* based on both rules *are dependent on sample size m*, in such a way which agrees with good sense.

But, of course, we do not know the true parameter θ, and thus $\mathfrak{p}$ and $\mathfrak{m}$ have to be estimated, by means of the sample that we already have, the actual k-size sample. We shall suggest two procedures based on a frequentist and a Bayesian solution, which shall be referred to hereafter as the $\mathfrak{f}$-*procedure* and $\mathfrak{b}$-*procedure*, respectively.

The frequentist solution consists in the construction of a $(1-\alpha)$-confidence interval for $\mathfrak{p}$ or $\mathfrak{m}$, of the form $[\mathfrak{p}_*,1]$ or $[\mathfrak{m}_*,\infty)$ based on the actual sample $x=(x_1,\ldots,x_k)\in\chi^k$, where $\mathfrak{p}_*$ and $\mathfrak{m}_*$ are estimates or underestimates of $\mathfrak{p}$ and $\mathfrak{m}$, respectively. Moreover, we have to choose a confidence interval in such a way that

$$\mathfrak{p}_*\xrightarrow{\mathscr{P}}\mathfrak{p}\quad\text{or}\quad\mathfrak{m}_*\xrightarrow{\mathscr{P}}\mathfrak{m}$$

as $k\to\infty$, where $\mathscr{P}$ denotes convergence in probability. This condition guarantees the appropriate behaviour of the $\mathfrak{f}$-procedure.

In many cases, $\mathfrak{p}$ and $\mathfrak{m}$ are increasing functions of $\rho^2(\Theta_0,\theta)$, therefore the desired confidence interval can be achieved by obtaining one for the last quantity of the form $[\rho_*^2,\infty)$, where ρ_*^2 is an estimate or underestimate of $\rho^2(\Theta_0,\theta)$. This strategy is convenient since, often, there exists an estimator of $\rho^2(\Theta_0,\theta)$ whose distribution depends on the parameters only through the single scalar quantity $\rho^2(\Theta_0,\theta)$, or at least, asymptotically. Then we can replace $\rho^2(\Theta_0,\theta)$ by ρ_*^2 in both the $\mathfrak{p}$ or $\mathfrak{m}$ rules.

Sometimes, in order to build confidence regions, it will be necessary to use asymptotic results. A useful one, for first-order efficient estimators $T_k(x)$ (such as maximum likelihood), is the following:

$$\sqrt{k}(\rho(T_k(x),\theta_0)-\rho(\theta,\theta_0))\xrightarrow{\mathscr{L}}Z\sim \mathrm{N}(0,1),$$

where θ is the parameter corresponding to true density function, $\theta_0\neq\theta$, $\mathscr{L}$ denotes convergence in law and $\mathrm{N}(0,1)$, the standard univariate normal distribution, as can easily be checked.

For the Bayesian solution, we shall assume a non-informative prior distribution, in particular one proportional to the Riemannian volume (although other prior distributions could be used, if there were strong reasons for so doing). We can take the average of the quantities $\mathfrak{p}$ or $\mathfrak{m}$ with respect to the corresponding posterior distribution, if additionally we have a sample $x\in\chi^k$. In any case, we have to ensure that this posterior

probability distribution converges weakly to the true probability measure, as $k \to \infty$, in order to ensure that the $\mathfrak{b}$-procedure behaves appropriately.

Although, following this procedure, we may use any convenient estimator, such as the maximum likelihood estimator, it is natural to consider now the (generalized) Bayes estimator taking as the (improper) prior distribution the Riemannian volume, and the square of the Rao distance as the loss function, see Oller and Corcuera (1995). Assuming certain additional conditions for Θ and the posterior measure, this estimator, which is the center of mass of the posterior distribution, exists and is uniquely defined, see Emery and Mokobodzki (1991) or Karcher (1977).

Thus, it would seem possible to combine the $\mathfrak{p}$ or $\mathfrak{m}$ rules with the $\mathfrak{f}$ and $\mathfrak{b}$ procedures to obtain four basically similar methods, replacing $\mathfrak{p}$ or $\mathfrak{m}$ by their underestimates $\mathfrak{p}_*$ or $\mathfrak{m}_*$ in the $\mathfrak{f}$-procedure and their average computed with the posterior distribution, in the $\mathfrak{b}$-procedure. Each of these methods is dependent on the selected estimator T_m. We shall refer to them hereafter as the $\mathfrak{pf}$, $\mathfrak{pb}$, $\mathfrak{mf}$, $\mathfrak{mb}$-methods and they will be developed in certain specific cases.

Notice also that in the $\mathfrak{f}$-procedure, the level α does not play the role of a *conditional probability of error* (as would the confidence level of a test), since it controls the estimates $\mathfrak{p}_*$ or $\mathfrak{m}_*$ for $\mathfrak{p}$ or $\mathfrak{m}$. Reasonably, α should be greater than 0 and less than or equal to 0.5, but probably it is not convenient to select values as low as 0.05 because smaller values for α imply lower values for $\mathfrak{p}_*$ and $\mathfrak{m}_*$, and thus we would obtain exaggerated underestimates for $\mathfrak{p}$ and $\mathfrak{m}$. In the following sections we illustrate these methods with several examples and compare the results obtained.

3. Examples

As is customary, since in the following examples the parameter space Θ is a subset of $\mathbb{R}^n$, we can identify the parameters with their coordinates. The whole parameter space will be denoted by Θ and the submanifold corresponding to the null hypothesis, by Θ_0.

Example 3.1. Let us consider the n-variate normal distribution with covariance matrix $\Sigma = \Sigma_0$ a known $n \times n$ positive definite symmetric matrix, $X \sim N_n(\mu, \Sigma_0)$, and the *testing problem* given by

$$\mathrm{H}_0 : H\mu + b = 0, \quad \mathrm{H}_1 : H\mu + b \neq 0,$$

where H is a $r \times n$ matrix with $0 < \operatorname{rank} H = r \leqslant n$ and b is a $r \times 1$ vector. Observe, in particular, the simple null hypothesis case

$$\mathrm{H}_0 : \mu = \mu_0, \quad \mathrm{H}_1 : \mu \neq \mu_0$$

obtained if we let $H = I$, $b = -\mu_0$, and then $r = n$.

The whole parameter space (model) will be $\Theta = \mathbb{R}^n$, and the null hypothesis corresponding parameter space (submodel) $\Theta_0 = \{\mu \in \mathbb{R}^n : H\mu + b = 0\}$ which is a $(n - r)$-dimensional linear manifold.

The *information metric* is given, in matrix notation, by

$$ds^2 = d\mu' \Sigma_0^{-1} d\mu,$$

the geometry is Euclidean, the Rao distance is in this case the *Mahalanobis distance*, whose square is

$$\rho^2(\mu_1, \mu_2) = (\mu_1 - \mu_2)' \Sigma_0^{-1} (\mu_1 - \mu_2)$$

and the Riemannian volume is

$$dV = (\det \Sigma_0)^{-1/2} d\mu.$$

Given a m-size sample $x = (x_1, \ldots, x_m) \in (\mathbb{R}^n)^m$, a natural estimator for this model is that of maximum-likelihood,

$$T_m(x) = \bar{x}_m = \frac{1}{m}(x_1 + \cdots + x_m).$$

Observe that this estimator is also the *centre of mass* of the posterior probability measure, based on a *non-informative prior* proportional to the Riemannian volume, which is, given a m-size sample $x = (x_1, \ldots, x_m) \in (\mathbb{R}^n)^m$,

$$\mu \sim N_n\left(\bar{x}_m, \frac{1}{m}\Sigma_0\right), \tag{3}$$

since in the Euclidean case, with a convenient parametrization, the centre of mass is just the expected value of X.

$\mathfrak{p}\mathfrak{f}$-*method*. First of all, given a k-size sample, $x = (x_1, \ldots, x_k) \in (\mathbb{R}^n)^k$, we have to determine an $(1-\alpha)$-confidence interval for $\mathfrak{p}$, of the form $[\mathfrak{p}_*, 1]$. In order to achieve this, notice that, if we let $\bar{X}_k \equiv T_k(\cdot)$, then

$$k\rho^2(\bar{X}_k, \mu) = k(\bar{X}_k - \mu)' \Sigma_0^{-1} (\bar{X}_k - \mu)$$

has an ordinary chi-square distribution with n-degrees of freedom. Therefore $\mathfrak{p}$ is a monotone increasing function of $\rho^2(\Theta_0, \mu)$, and thus we can obtain the above-mentioned confidence interval for $\mathfrak{p}$ through a confidence interval for $\rho^2(\Theta_0, \mu)$ of the form $[\rho_*^2, \infty)$, being ρ_*^2 an underestimate of $\rho^2(\Theta_0, \mu)$.

To obtain this interval, notice that the geometry corresponding to the model is Euclidean, and therefore the squared distance between $\eta \in \Theta$ to Θ_0 is given by

$$\rho^2(\Theta_0, \eta) = (\eta + H^- b)' H' (H \Sigma_0 H')^{-1} H (\eta + H^- b)$$

independently of the g-inverse, H^-, choice. Thus, the square of the distance between the true μ value, and the estimated one from $\bar{x}_k$, to Θ_0 will be given, by applying this formula, by $\rho^2(\Theta_0, \mu)$ and $\rho^2(\Theta_0, \bar{x}_k)$, respectively.

Moreover, observe that $W = \Sigma_0^{1/2} H' (H \Sigma_0 H')^{-1} H \Sigma_0^{1/2}$ is the projector operator, according to the information metric, into the orthogonal subspace of $\{\eta \in \mathbb{R}^n : H\eta = 0\}$, the director subspace corresponding to Θ_0, consequently is a symmetric rank r idempotent

matrix, and taking into account that the statistic

$$\sqrt{k}\Sigma_0^{-1/2}(\bar{X}_k + H^- b) \sim N_n(\sqrt{k}\Sigma_0^{-1/2}(\mu + H^- b), I)$$

it follows that

$$\begin{aligned} k\rho^2(\Theta_0, \bar{X}_k) &= k(\bar{X}_k + H^- b)' H'(H\Sigma_0 H')^{-1} H(\bar{X}_k + H^- b) \\ &= k(\bar{X}_k + H^- b)' \Sigma_0^{-1/2} W \Sigma_0^{-1/2}(\bar{X}_k + H^- b) \end{aligned}$$

has a non-central chi-square distribution $\chi_r^2(\delta)$ with r degrees of freedom and noncentrality parameter $\delta = k\rho^2(\Theta_0, \mu)$, see Muirhead (1982, pp. 26–32).

Thus, if we define $\gamma_{r,\alpha}(\delta)$ through

$$P(\chi_r^2(\delta) \leqslant \gamma_{r,\alpha}(\delta)) = 1 - \alpha,$$

where $\gamma_{r,\alpha}(\delta)$ is, clearly, a strictly monotone increasing function of δ, and given $\rho^2(\Theta_0, \bar{x}_k)$, the $(1-\alpha)$-confidence interval for $\rho^2(\Theta_0, \mu)$ will be $[\rho_*^2, \infty)$ with

$$\rho_*^2 = \begin{cases} \dfrac{\gamma_{r,\alpha}^{-1}(k\rho^2(\Theta_0, \bar{x}_k))}{k} & \text{if } k\rho^2(\Theta_0, \bar{x}_k) > \gamma_{r,\alpha}(0), \\ 0 & \text{otherwise.} \end{cases}$$

Then we have to evaluate $\mathfrak{p}_* = 2P(\rho^2(\bar{X}_m, \mu) < \rho_*^2) - 1$, through the χ_n^2 distribution tables and finally we shall accept H_0 if $\mathfrak{p}_*$, an underestimate of $\mathfrak{p}$, is less than zero. Observe that if we let ξ_n be the median of a χ_n^2 distribution, we shall reject the submodel if

$$m\rho_*^2 = \frac{m}{k}\gamma_{r,\alpha}^{-1}(k\rho^2(\Theta_0, \bar{x}_k)) > \xi_n$$

accepting it otherwise.

Notice also that if we compare this result with the classical, and natural, hypothesis test which rejects the null hypothesis, H_0, for large values of $k\rho^2(\Theta_0, \bar{X}_k)$, if we *accept* H_0 with a *significance level* α, then we shall also *accept* the null hypothesis with the $\mathfrak{pf}$-method, though for *different* reasons, since in this case $\rho_*^2 = 0$ and thus $P(\rho^2(\bar{X}_m, \mu) < \rho_*^2) = 0$, and $\mathfrak{p}_* = -1 \leqslant 0$. But, if with the classical test, we *reject* H_0, with the $\mathfrak{pf}$-method we shall also reject it if

$$2P(\chi_n^2 < m\rho_*^2) - 1 > 0$$

and we still *accept* H_0 if the above-mentioned quantity is less than zero, which depends on the value of m: small values allow the acceptance of the submodel.

$\mathfrak{mf}$-*method.* If $k\ \rho^2(\Theta_0, \bar{x}_k) > \gamma_{r,\alpha}(0)$, expression (2) becomes

$$m\rho_*^2 - n = \frac{m}{k}\gamma_{r,\alpha}^{-1}(k\rho^2(\Theta_0, \bar{x}_k)) - n > 0 \tag{4}$$

since $E(\chi_n^2) = n$, rejecting the null hypothesis if (4) occurs. In the other case, $k\ \rho^2(\Theta_0, \bar{x}_k) \leqslant \gamma_{r,\alpha}(0)$, we always accept the null hypothesis.

$\mathfrak{pb}$-*method.* Observe now that we have to take the average of the quantity

$$P(\rho^2(T_m, \mu) < \rho^2(\Theta_0, \mu))$$

with respect to the posterior distribution (3), but now for a k size sample: $\mu \sim N_n(\bar{x}_k, (1/k)\Sigma_0)$.

We have to take into account that given $\bar{x}_k$, with the same basic notation as in the previous sections,

$$\sqrt{k}\Sigma_0^{-1/2}(\mu + H^- b) \sim N_n(\sqrt{k}\Sigma_0^{-1/2}(\bar{x}_k + H^- b), I)$$

and thus,

$$k\rho^2(\Theta_0, \mu) = k(\mu + H^- b)'\Sigma_0^{-1/2} W \Sigma_0^{-1/2}(\mu + H^- b)$$

has a non-central chi-square distribution $\chi_r^2(\delta)$ with r degrees of freedom and non-centrality parameter $\delta = k\rho^2(\Theta_0, \bar{x}_k)$. Moreover, $m\rho^2(T_m, \mu) \sim \chi_n^2$, *independently* of μ. Therefore

$$E(P(\rho^2(T_m, \mu) < \rho^2(\Theta_0, \mu))|\bar{x}_k) = P\left(\frac{kn}{mr} < Y\right),$$

where

$$Y = \frac{k\rho^2(\Theta_0, \mu)/r}{m\rho^2(T_m, \mu)/n}$$

is a random variable with a $F_{r,n}(\delta)$ distribution, i.e., a non-central F distribution, with r and n degrees of freedom and non-centrality parameter $\delta = k\rho^2(\Theta_0, \bar{x}_k)$.

If this probability is greater than $\frac{1}{2}$ then we shall *reject* the null hypothesis (*accept* the whole model), and vice versa. Observe that if we let $\zeta_{r,n}(\delta)$ be the median of a $F_{r,n}(\delta)$ distribution, we shall reject the submodel if

$$\zeta_{r,n}(\delta) > \frac{kn}{mr}$$

accepting it otherwise. Notice also that if we let $m \to \infty$ then we shall certainly reject the null hypothesis since $\zeta_{r,n}(\delta) > 0$.

mb-*method*. The mean square Rao distances for T_m is given by

$$E_\mu(\rho^2(T_m, \mu)) = E_\mu((\bar{X}_m - \mu)'\Sigma_0^{-1}(\bar{X}_m - \mu)) = \frac{n}{m}$$

and under the null hypothesis the error will be measured as $\rho^2(\Theta_0, \mu)$.

Now we shall have to compute the average of both quantities with respect to the posterior distribution: $\mu \sim N_n(\bar{x}_k, (1/k)\Sigma_0)$, we shall have

$$E(E_\mu(\rho^2(T_m, \mu))\,|\,\bar{x}_k) = \frac{n}{m}$$

and since given $\bar{x}_k$, $k\rho^2(\Theta_0, \mu)$ has a non-central chi-square distribution $\chi_r^2(\delta)$ with r degrees of freedom and non-centrality parameter $\delta = k\rho^2(\Theta_0, \bar{x}_k)$, with $E(\chi_r^2(\delta)) = r + \delta$, we have

$$E(\rho^2(\Theta_0, \mu)\,|\,\bar{x}_k) = \frac{r}{k} + \rho^2(\Theta_0, \bar{x}_k).$$

Therefore, if

$$\frac{r}{k} + \rho^2(\Theta_0, \bar{x}_k) - \frac{n}{m} > 0$$

Table 1
The considered methods summarized for the n-variate normal model

Method	Condition
$\mathfrak{pf}$	$m\rho_*^2 > \xi_n$
$\mathfrak{mf}$	$m\rho_*^2 > n$
$\mathfrak{pb}$	$m\frac{r}{k}\zeta_{r,n}(\delta) > n$
$\mathfrak{mb}$	$m(\frac{r}{k} + \rho^2(\Theta_0, \bar{x}_k)) > n$

it is advantageous (from the present point of view) to *reject* H_0, or we shall select the whole model $X \sim N_n(\mu, \Sigma_0)$, $\mu \in \mathbb{R}^n$. Otherwise, we shall *accept* the null hypothesis (submodel) $X \sim N_n(\mu, \Sigma_0)$, $\mu \in \Theta_0$.

Equivalently, if we are going to work with samples of size m and

$$m \leqslant \frac{nk}{r + k\rho^2(\Theta_0, \bar{x}_k)}$$

it is *preferable to work with the submodel* $X \sim N_n(\mu, \Sigma_0)$, $\mu \in \Theta_0$, since in this case we shall not have enough information to make good estimates with the whole model, because under this, the expected loss will be bigger than with the submodel. Notice that for large enough sample sizes this upper bound for m is approximately equal to $n/\rho^2(\Theta_0, \bar{x}_k)$.

Moreover, observe that if

$$\frac{nk}{r + k\rho^2(\Theta_0, \bar{x}_k)} < 1 \quad \text{or equivalently} \quad \frac{nk - r}{k} < \rho^2(\Theta_0, \bar{x}_k)$$

then the null hypothesis H_0 will *always* be rejected, independently of the sample size m.

In summary, a comparison might be made of the conditions for the *rejection* of the null hypothesis (submodel) in the various methods described.

From Table 1, using normal approximations for χ_n^2, $\chi_r^2(\delta)$ and $F_{r,n}(\delta)$ distributions, given in Abramowitz (1970, pp. 942–948), we obtain, for large n,

$$\xi_n \sim n\left(1 - \frac{2}{9n}\right)^3 \sim n - \frac{2}{3}$$

and

$$\zeta_{r,n}(\delta) \sim \left(1 + \frac{k}{r}\rho^2(\Theta_0, \bar{x}_k)\right)\left(\frac{9n}{9n-2}\left(1 - \frac{2(r + 2k\rho^2(\Theta_0, \bar{x}_k))}{9(r + k\rho^2(\Theta_0, \bar{x}_k))^2}\right)\right)^3$$

then

$$\frac{r}{k}\zeta_{r,n}(\delta) \sim \beta_{rn}\left(\frac{r}{k} + \rho^2(\Theta_0, \bar{x}_k)\right)$$

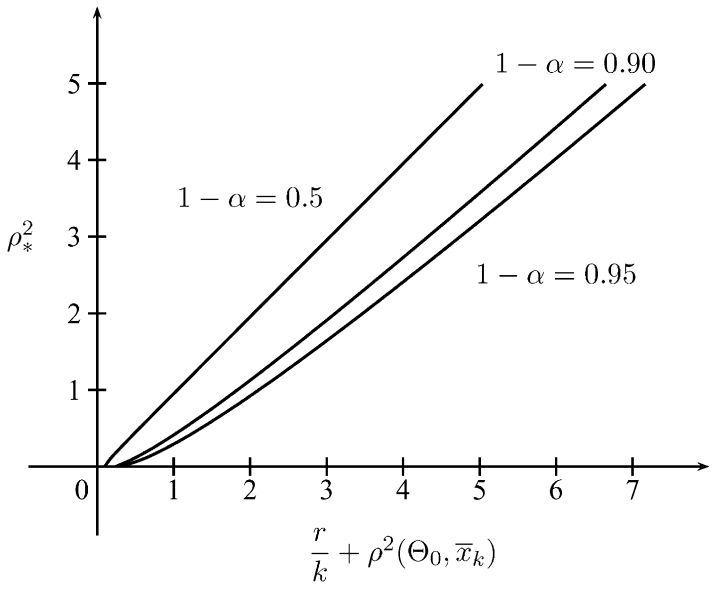

Fig. 1. Graphic output obtained with $r=1$, $k=15$.

where

$$\beta_{rn}\sim\left(\frac{n(9r-2)}{(9n-2)r}\right)^3 \quad \text{or} \quad \beta_{rn}\sim\left(\frac{9n}{9n-2}\right)^3$$

for small values of $\rho^2(\Theta_0,\bar{x}_k)$) or for large values, respectively.

Therefore, on the one hand $\mathfrak{pf}$ and $\mathfrak{mf}$, will supply similar conclusions since $\xi_n\sim n-\frac{2}{3}$, on the other hand, $\mathfrak{pb}$ and $\mathfrak{mb}$ behave similarly since, approximately, $(r/k)\zeta_{r,n}(\delta)$ is proportional to $r/k+\rho^2(\Theta_0,\bar{x}_k)$.

In order to compare $\mathfrak{f}$ and $\mathfrak{b}$ procedures, we may plot Figs. 1 and 2, $r/k+\rho^2(\Theta_0,\bar{x}_k)$ versus ρ_*^2 for different values of r and k, and considering also different $(1-\alpha)$ confidence intervals for ρ_*^2. To obtain these plots we have used a normal approximation for a non-central chi-square distribution: if $U\sim\chi_r^2(\delta)$, then the random variable

$$\sqrt{\frac{9(r+2\delta)^2}{2(r+2\delta)}}\left(\left(\frac{U}{r+\delta}\right)^{1/3}+\frac{2(r+2\delta)}{9(r+2\delta)^2}-1\right)$$

has, approximately, a standard univariate normal distribution.

Observe, first of all, that the plot does not appreciably depend on particular r and k values. From both graphs, it is clear that both $\mathfrak{f}$- and $\mathfrak{b}$-procedures behave in a similar way, *provided* $r/k+\rho^2(\Theta_0,\bar{x}_k)$ is greater than a certain quantity, c_α, which depends on the confidence level $(1-\alpha)$. For confidence levels between 0.50 and 0.95, c_α oscillates between 0.098 and 0.535. For smaller values of $r/k+\rho^2(\Theta_0,\bar{x}_k)$, the $\mathfrak{f}$- and $\mathfrak{b}$-procedures differ considerably: the $\mathfrak{f}$-procedure gives priority to the submodel (null hypothesis) over the whole, compared with the $\mathfrak{b}$-procedure, a fact which indicates that

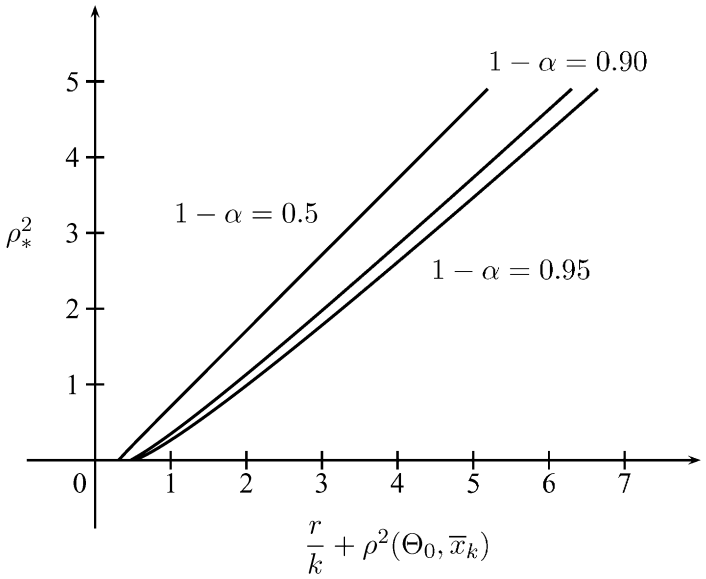

Fig. 2. Graphic output obtained with $r=5$, $k=30$.

the ƒ is more robust than ƀ-procedure with respect to the departure from the basic assumption that the unknown prior distribution is dominated by the Riemannian volume.

Example 3.2. Let us consider a linear normal model $Y \sim N_n(X\beta, \sigma^2 I)$, where $\beta \in \mathbb{R}^s$ and $\sigma > 0$, are unknown parameters and X is a $n \times s$ matrix whose coefficients are known and such that $0 < \operatorname{rank} X = s \leqslant n$, with the *testing problem*:

$$\mathrm{H}_0 : H\beta = 0, \quad \mathrm{H}_1 : H\beta \neq 0,$$

where H is a $r \times s$ matrix with $0 < \operatorname{rank} H = r \leqslant s$. *Note*: the general case: $\mathrm{H}_0 : H\beta = c$, $\mathrm{H}_1 : H\beta \neq c$, can obviously be reduced to the first one through convenient affine changes in the parameters and variables.

Hereafter, in this example, the elements of $\mathbb{R}^k$ are considered, in matrix notation, as column vectors. Under this parameterization the whole parameter space is $\Theta = \mathbb{R}^s \times \mathbb{R}^+$ and the parameter space corresponding to the null hypothesis $\Theta_0 = \{(\beta, \sigma) \in \mathbb{R}^s \times \mathbb{R}^+ : H\beta = 0\}$, with $\operatorname{rank} H = r$.

The *information metric* is given, in matrix notation, by

$$\mathrm{d}s^2 = \frac{1}{\sigma^2}(\mathrm{d}\beta' X' X \, \mathrm{d}\beta + 2n \, \mathrm{d}\sigma^2),$$

which is the *Poincaré hyperbolic metric* of the upper half space $\mathbb{R}^s \times \mathbb{R}^+$, see Burbea and Oller (1988), and the Riemannian distance is

$$\rho(\beta_1, \sigma_1; \beta_2, \sigma_2) = \sqrt{2n} \log \frac{1+\delta_{12}}{1-\delta_{12}} = 2\sqrt{2n} \tanh^{-1}(\delta_{12}),$$

where

$$\delta_{12} = \sqrt{\frac{\|\beta_2 - \beta_1\|^2 + 2n(\sigma_1 - \sigma_2)^2}{\|\beta_2 - \beta_1\|^2 + 2n(\sigma_1 + \sigma_2)^2}}$$

with

$$\|\beta_2 - \beta_1\|^2 = (\beta_1 - \beta_2)'X'X(\beta_1 - \beta_2).$$

For this example we are going to develop the pf-method, considering the maximum-likelihood estimator. Let $\tilde{X}_{(m)}$ be an $nm \times s$ matrix defined as

$$\tilde{X}_{(m)} = 1_m \otimes X,$$

where $1'_m = (1, \overset{m)}{\ldots}, 1)'$, where $\otimes$ stands for the Kronecker product. Therefore, the random $n \times m$ matrix $Y = (Y_1, \ldots, Y_m)$, which columns are random vectors corresponding to a linear model sample of size m, can be considered a 1-size sample of the linear model $\mathrm{vec}(Y) \sim N_{nm}(\tilde{X}_{(m)}\beta, \sigma^2 I)$, and thus, taking into account that

$$(1_m \otimes X)'1_m \otimes X = mX'X,$$

$$(1_m \otimes X)'\mathrm{vec}(Y) = mX'\bar{Y}_m$$

and

$$\mathrm{vec}(YH_m)'\mathrm{vec}(YH_m) = \sum_{i=1}^{m}(Y_i - \bar{Y}_m)'(Y_i - \bar{Y}_m) = \mathrm{tr}(YH_mY')$$

where $\bar{Y}_m = (1/m)(Y_1 + \cdots + Y_m)$ and H_m is the $m \times m$ symmetric idempotent matrix $H_m = I - (1/m)1_m1'_m$, then the maximum-likelihood estimators will be given by

$$\beta^*_{(m)} = (\tilde{X}'_{(m)}\tilde{X}_{(m)})^{-1}\tilde{X}'_{(m)}\,\mathrm{vec}(Y) = (X'X)^{-1}X'\bar{Y}_m$$

and

$$\sigma^*_{(m)} = \sqrt{\frac{1}{nm}\,\mathrm{vec}(Y)'(I - \tilde{X}_{(m)}(\tilde{X}'_{(m)}\tilde{X}_{(m)})^{-1}\tilde{X}'_{(m)})\mathrm{vec}(Y)}$$

$$= \sqrt{\frac{1}{nm}\,\mathrm{tr}(YH_mY') + \frac{1}{n}\,\bar{Y}'_m(I - X(X'X)^{-1}X')\bar{Y}_m}$$

Clearly $\bar{Y}_m \sim N_n(X\beta, \frac{\sigma^2}{m}I)$, then

$$\frac{m}{\sigma^2}\,\bar{Y}'_m(I - X(X'X)^{-1}X')\bar{Y}_m \sim \chi^2_{n-s}$$

and

$$\frac{1}{\sigma^2}\,\mathrm{tr}(YH_mY') = \frac{1}{\sigma^2}\sum_{j=1}^{n}\sum_{i=1}^{m}(Y_{ij} - \bar{Y}_{\cdot j})^2 \sim \chi^2_{nm-n}.$$

Therefore, if we define:

$$U_{(m)} = m\|\beta^*_{(m)} - \beta\|^2/\sigma^2 \sim \chi^2_s \quad V_{(m)} = nm(\sigma^*_{(m)})^2/\sigma^2 \sim \chi^2_{nm-s}$$

with $U_{(m)}$ and $V_{(m)}$ independent, we have

$$\begin{aligned} &P(\rho^2(\beta^*_{(m)}, \sigma^*_{(m)}; \beta, \sigma) < c) \\ &= P\left(U_{(m)} + 2nm\left(\sqrt{\frac{V_{(m)}}{nm}} - 1\right)^2 < 8\sqrt{nmV_{(m)}}\sinh^2\left(\sqrt{\frac{c}{8n}}\right)\right) \\ &= \alpha_{s,nm}\int_a^b \int_0^{4\sqrt{nmv}\cosh(\sqrt{c/2n})-2v-2nm} u^{(s/2)-1}v^{(nm-s)/2-1}\mathrm{e}^{-(1/2)(u+v)}\,\mathrm{d}u\,\mathrm{d}v, \end{aligned} \tag{5}$$

where $c > 0, a = nm\mathrm{e}^{-\sqrt{2c/n}}$, $b = nm\mathrm{e}^{\sqrt{2c/n}}$ and

$$\alpha_{s,nm} = \frac{1}{2^{nm/2}\Gamma(s/2)\Gamma((nm-s)/2)}.$$

It is clear that this probability is a monotonous increasing function of c, and therefore $\mathfrak{p}$ is a monotone increasing function of $\rho^2(\Theta_0, (\beta, \sigma))$, thus, given a $n \times k$ real matrix $y = (y_1, \ldots, y_k)$ which columns correspond to a k-size sample of the linear model, we can obtain an $(1-\alpha)$-confidence interval for $\mathfrak{p}$, of the form $[\mathfrak{p}_*, 1]$, through a confidence interval for $\rho^2(\Theta_0, (\beta, \sigma))$ of the form $[\rho^2_*, \infty)$, being ρ^2_* an underestimate of $\rho^2(\Theta_0, (\beta, \sigma))$ given a k-size sample.

To obtain this interval, observe, after some straightforward computation, that

$$\rho(\Theta_0, (\beta, \sigma)) = 2\sqrt{2n}\tanh^{-1}\left(\frac{\sqrt{D^2+1}-1}{D}\right),$$

where

$$D^2 = D^2(\beta, \sigma) = \frac{\beta' H'(H(X'X)^{-1}H')^{-1}H\beta}{2n\sigma^2}$$

D^2 being independent of the g-inverse choice. Thus, $\rho(\Theta_0, (\beta, \sigma))$ is a monotonous increasing function of D^2, and a confidence interval for this quantity of the form $[\rho^2_*, \infty)$ can be obtained through a confidence interval for D^2 of the form $[D^2_*, \infty)$. Notice now that

$$\mathscr{Y} = \frac{2(nk-s)}{r} D^2(\beta^*_{(k)}, \sigma^*_{(k)})$$

is a random variable with a $F_{r,nk-s}(\Delta)$ distribution, with a non-centrality parameter $\Delta = 2nkD^2(\beta, \sigma)$. Let us define $\phi_{r,nk-s,\alpha}(\Delta)$ through

$$P(\mathscr{Y} \leqslant \phi_{r,nk-s,\alpha}(\Delta)) = 1 - \alpha,$$

where $\phi_{r,nk-s,\alpha}(\Delta)$ is a strictly monotone increasing function of Δ, therefore given $\hat{D}^2_{(k)} = D^2(\hat{\beta}^*_{(k)}, \hat{\sigma}^*_{(k)})$, where $\hat{\beta}^*_{(k)}$ and $\hat{\sigma}^*_{(k)}$ are the particular estimates corresponding to the k-size sample y, if we let $\delta \equiv \phi_{r,nk-s,\alpha}(0)$, a $(1-\alpha)$-confidence interval for D^2 will be $[D^2_*, \infty)$ with

$$D^2_* = \begin{cases} \dfrac{1}{2nk}\,\phi^{-1}_{r,nk-s,\alpha}\left(\dfrac{2(nk-s)}{r}\hat{D}^2_{(k)}\right) & \text{if } \dfrac{2(nk-s)}{r}\,\hat{D}^2_{(k)} > \delta, \\ 0 & \text{otherwise,} \end{cases}$$

and thus

$$\rho_* = 2\sqrt{2n}\tanh^{-1}\left(\frac{\sqrt{D^2_*+1}-1}{D_*}\right).$$

Next, we have to evaluate $\mathfrak{p}_* = 2P(m\rho^2(\beta^*_{(m)}, \sigma^*_{(m)}; \beta, \sigma) < m\rho^2_*) - 1$, through the distribution given in (5). As a matter of fact, we only need to know the median of the $m\rho^2(\beta^*_{(m)}, \sigma^*_{(m)}; \beta, \sigma)$ distribution. First of all, observe that

$$8nm\sqrt{\frac{V_{(m)}}{nm}}\sinh^2\left(\sqrt{\frac{c}{8nm}}\right) = 8nm\sqrt{\frac{V_{(m)}}{nm}}\,\frac{1}{2}\left(\cosh\left(\sqrt{\frac{c}{2nm}}\right)-1\right)$$

$$= \sqrt{\frac{V_{(m)}}{nm}}\left(c + O\left(\frac{1}{nm}\right)\right)$$

which follows through a second-order Taylor expansion of function cosh. Additionally, when $nm \to \infty$, we can easily prove that

$$\frac{V_{(m)} - nm}{\sqrt{2nm}} \xrightarrow{\mathscr{L}} Z \sim N(0,1), \quad \sqrt{2nm}\left(\sqrt{\frac{V_{(m)}}{nm}} - 1\right) \xrightarrow{\mathscr{L}} Z \sim N(0,1),$$

$$\frac{V_{(m)}}{nm} \xrightarrow{\mathscr{P}} 1.$$

Therefore,

$$8nm\sqrt{\frac{V_{(m)}}{nm}}\sinh^2\left(\sqrt{\frac{c}{8nm}}\right) = c + O_P\left(\frac{1}{nm}\right)$$

Finally, with fixed s, we conclude that

$$m\rho^2(\beta^*_{(m)}, \sigma^*_{(m)}; \beta, \sigma) \xrightarrow{\mathscr{L}} \chi^2_{s+1},$$

i.e., the limiting distribution of $m\rho^2(\beta^*_{(m)}, \sigma^*_{(m)}; \beta, \sigma)$ is chi-square with $s+1$ degrees of freedom.

Thus, for large values of nm, if we let ξ_{s+1} be the median of a χ^2_{s+1} distribution, we shall reject the submodel if $m\rho^2_* > \xi_{s+1} \sim s + \frac{1}{3}$ accepting it otherwise. For small values of nm we have computed the median numerically, obtaining Table 2.

Table 2
Median of $m\rho^2(\beta^*_{(m)}, \sigma^*_{(m)}; \beta, \sigma)$

nm	s				
	1	2	3	4	5
10	1.586	3.022	4.871	7.311	10.638
15	1.513	2.763	4.224	5.942	7.982
20	1.479	2.650	3.964	5.438	7.098
25	1.459	2.588	3.824	5.177	6.658
30	1.447	2.548	3.737	5.017	6.395
50	1.422	2.471	3.574	4.726	5.928
χ^2_{s+1}	1.386	2.366	3.357	4.351	5.348

4. Discussion

It has been shown that in a wide range of hypothesis testing problems, the classical approach based on the work of Neyman and Pearson (1933) give rise to several *epistemological* problems. From the authors' point of view, its value is only the value of a *conventional way of reasoning*. Therefore, within the framework outlined, it would seem convenient to clarify the meaning of the acceptance of the null hypothesis, since we *know* that it is false. In order to achieve this, several similar methods have been proposed and compared with classical techniques.

Broadly speaking, the newly proposed methods share four main features. First of all, they consider a hypothesis testing problem as one in which a choice has to be made between two models, one nested inside the other, in order to fit the data. Secondly, they rely on a convenient estimator for the whole model. Additionally, the methods are based on two samples: an *actual* sample, already obtained, and a *potential* sample. The former provides the available information, while the latter fixes the *resolving power* to be used in the future, where the final decision is dependent on both sample sizes. Finally, they try to evaluate the error corresponding to each model by adopting different approaches: frequentist or Bayesian and taking into account moments or probabilities. Here, we emphasize that the ƥ-rules are less dependent than ɱ-rules on the distance employed, and ƒ-procedures have an advantage over ɓ-procedures in that they do not depend on a particular prior distribution.

These methods allow us to choose the submodel (null hypothesis) instead of the model, in the belief that it is a better model to describe the observed phenomena, once the *resolving power*, by means of which we are observing the reality, has been fixed. These considerations suggest the possibility of developing a formal structure for a certain class of inductive reasoning, based on a measure of *closeness* to *the truth* rather than on the *probability of error*, since *the truth* is not something that *we possess* but rather something that *we are for ever seeking*, as in life itself!

To conclude this section, recently, we have become aware of some interesting concepts in the field of the design of experiments, introduced in the previous decades, which are closely related with our work. Srivastava (1984) introduced the concept of

revealing power of a design, which, broadly speaking, refers to the ability of a design to reveal or identify the correct model, and he applies this in Srivastava (1992) and Srivastava (1987). Moreover, several measures of this concept for the case of sequential design, are introduced in Srivastava and Hveberg (1992).

The *revealing power* is similar to the notion, not defined explicitly in our paper, of *resolving power*, which we use in the sense of the ability, using the information contained in a sample, to discriminate different probability measures as really different.

Other useful references for modelling are Srivastava (1996), with an interesting methodological discussion and Srivastava (1990) with a modern overview of factor design theory.

References

Abramowitz, M., 1970. Handbook of Mathematical Functions. Dover, New York.

Amari, S., 1985. Differential-Geometrical Methods in Statistics. Vol. 28. Springer, New York.

Atkinson, C., Mitchell, A.F.S., 1981. Rao's distance measure. Sankhyà 43 A, 345–365.

Berger, J.O., 1985. Statistical Decision Theory. Springer, New York.

Burbea, J., 1986. Informative geometry of probability spaces. Expositiones Mathematicae 4, 347–378.

Burbea, J., Oller, J.M., 1988. The information metric for univariate linear elliptic models. Statist. Decisions 6, 209–221.

Burbea, J., Rao, C.R., 1982. Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. J. Multivariate Anal. 12, 575–596.

Čencov, N.N., 1982. Statistical Decision Rules and Optimal Inference, Trans. Math. Monographs, Vol. 53, American Mathematical Society, Providence, RI (English translation of the Russian book published in 1972, Nauka, Moscow).

Emery, M., Mokobodzki, G., 1991. Sur le barycentre d'une probabilité dans une variété, in: Séminaire de Probabilités XXV: Lecture Notes in Mathematics, Vol. 1485. Springer, Berlin, pp. 220–233.

Good, I.J., 1988. Scientific method and statistics. In: Kotz, S., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 8. Wiley, New York, pp. 291–304.

Harlow, L.L., Mulaik, S.A., Steiger, J.H., 1997. What if there were no Significance Tests. Lawrence Erlbaum Associates, Mahwah, NJ.

Jeffreys, H., 1946. An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. 186 A, 453–461.

Karcher, H., 1977. Riemannian center of mass and mollifier smoothing. Comm. Pure Appl. Math. 30, 509–541.

Muirhead, R.J., 1982. Aspects of Multivariate Statistical Theory. Wiley, New York.

Neyman, J., Pearson, E.S., 1933. On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. Roy. Soc. 231 A, 289–337.

Oller, J.M., Corcuera, J.M., 1995. Intrinsic analysis of statistical estimation. Ann. Statist. 23 (5), 1562–1581.

Rao, C.R., 1945. Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. 37, 81–91.

Rollal, R., 1997. Statistical Evidence, a Likelihood Paradigm. Chapman & Hall, Boca Raton, FL.

Srivastava, J.N., 1984. Sensitivity and revealing power: two fundamental statistical criteria other than optimality arising in discrete experimentation. In: Hinkelmann, K. (Ed.), Experimental Designs, Statistical Models, and Genetic Statistics Models, and Genetic Statistics. Marcel Dekker, New York, pp. 95–117.

Srivastava, J.N., 1987. On the inadequacy of the customary orthogonal arrays in quality control and scientific experimentation, and the need of probing designs. Comm. Statist. 16, 2901–2941.

Srivastava, J.N., 1990. Modern factorial design theory for experimenters. In: Gosh, S. (Ed.), Design and Analysis of Industrial Experiments. Marcel Dekker, New York, pp. 311–406.

Srivastava, J.N., 1992. A 2^8 factorial search design with good revealing power. Sankhya 54, 461–474.

Srivastava, J.N., 1996. A critique of some aspects of experimental design. In: Gosh, S., Rao, C.R. (Eds.), Handbook of Statistics, Design and Analysis of Experiments, Vol. 13. Elsevier, North-Holland, New York, pp. 309–341.

Srivastava, J.N., Hveberg, R., 1992. Sequential factorial probing designs for identifying and estimating nonnegligible parameters. J. Statist. Plann. Inference 30, 141–162.