Logical consistency in simultaneous statistical test procedures

RAFAEL IZBICKI*, *Statistics Department, Federal University of São Carlos, Brazil.*

LUÍS GUSTAVO ESTEVES**, *Statistics Department, Universidade de São Paulo, Brazil.*

Abstract

Many authors have argued that, when performing simultaneous statistical test procedures, one should seek for solutions that lead to decisions that are consistent and, consequently, easier to communicate to practitioners of statistical methods. In this way, the set of hypotheses that are rejected and the set of hypotheses that are *not* rejected by a testing procedure should be consistent from a logic standpoint. For instance, if hypothesis A implies hypothesis B, a procedure that rejects B should also reject A, a property not always met by multiple test procedures. We contribute to this discussion by exploring how far one can go in constructing coherent procedures while still preserving statistical optimality. This is done by studying four types of logical consistency relations. We show that although the only procedures that satisfy more than (any) two of these properties are simple tests based on point estimation, it is possible to construct various interesting methods that fulfil one or two of them while preserving different statistical optimality criteria. This is illustrated with several Bayesian and frequentist examples. We also characterize some of these properties under a decision-theoretic framework.

Keywords: Simultaneous test procedures, hypotheses testing, logical coherence, coherence principle, consonance principle.

1 Introduction

In many scientific problems, one is interested in testing several hypotheses simultaneously. Such a situation is called a *multiple* (or *simultaneous*) *hypotheses testing* problem in statistics literature [45]. This is typical, e.g. in clinical trials where one is interested in comparing the effectiveness of drugs and their side effects, or in genetic experiments involving microarrays. See more examples in [13].

Many times, decisions resulting from multiple hypotheses tests may lead to epistemic confusions because of inconsistencies between the hypotheses that are rejected and those that are not rejected on the basis of such tests. We refer to these shortcomings as logical inconsistencies because the origin of the confusion is the lack of logical coherence between hypotheses rejected and hypotheses not rejected: if the results of the tests are considered to be degenerate truth values (0 or 1), it may be seen as a logical issue. For instance, this is shown by [15] in a regression setting. Considering the linear model $E[Y|x]=\beta_0+\beta_1 x+\beta_2 x^2$, they show that the Bonferroni–Holm testing procedure can reject $\beta_2=0$, but not reject $\beta_1=\beta_2=0$. As $\beta_1=\beta_2=0$ implies $\beta_2=0$ from a logic standpoint, these conclusions may be confusing for a practitioner. Indeed, on the grounds that the test for $\beta_2=0$ rejects it and that $\beta_1=\beta_2=0$ implies $\beta_2=0$, a decision-maker can decide to reject $\beta_1=\beta_2=0$. On the other hand, he can analogously decide not to reject $\beta_2=0$ from the fact that the test for $\beta_1=\beta_2=0$ does not reject it! The fact that while $\beta_2=0$ is rejected by Bonferroni–Holm testing procedure

*E-mail: rizbicki@ufscar.br
**E-mail: lesteves@ime.usp.br

Vol. 23 No. 5, © The Author 2015. Published by Oxford University Press. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
doi:10.1093/jigpal/jzv027 Advance Access published 24 June 2015

$\beta_1 = \beta_2 = 0$ is not—even though from a logic perspective $\beta_1 = \beta_2 = 0$ implies $\beta_2 = 0$—leads to an epistemic confusion that makes it hard to report the results obtained by the testing procedure, and which is also many times embarrassing [38, 52, 54]. As another example, [42] describes the same incoherence in the case *E.E.O.C. Federal Reserve Bank of Richmond* [40] '*In this lively exchange the plaintiff's statistical experts tries to explain to a judge why one should use a one-sided test (with P value 0.037 in this example) rather than a two-sided test (with P value 0.074). The significance of the choice of the hypothesis was quite apparent to the judge*.' The problem in this example is that while the two-sided hypothesis ($\mu = 0$) is not rejected at the 5% level, the one-sided ($\mu \leq 0$) is. However, $\mu = 0$ implies $\mu \leq 0$. See also [35] for an interesting example where, in an Analysis of variance (ANOVA) setting, likelihood ratio tests for hypotheses $\mu_1 = \mu_2$ and $\mu_1 = \mu_2 = 0$ lead to rejection of the former, but not rejection of the latter.

Some authors therefore argue that some times one should waive on maximizing standard efficiency criteria (such as power of the tests) to produce coherent results that are easier to communicate to non-statisticians: '*One could … argue that 'power is not everything.' In particular for multiple test procedures one can formulate additional requirements, such as, for example, that the decision patterns should be logical, conceivable to other persons, and, as far as possible, simple to communicate to non-statisticians.'* [15].

How far can one go in constructing coherent procedures while still preserving statistical optimality? In this work, we try to answer this question by providing a framework for evaluating logical coherence in simultaneous test procedures.

1.1 Background

Several methods aim at creating optimal statistical tests for simultaneous procedures. From a Bayesian point of view, most approaches consist in minimizing (posterior) expected loss functions for the hypotheses of interest (e.g. [7, 13, 53]). From a frequentist perspective, various criteria have been introduced. Among them, popular approaches are controlling the error rate per family (PFE), the family wise error rate (FWER) and the false discovery rate (FDR) [2, 10, 45]. The reader is referred to [13], [45] and [9] for a review on simultaneous tests procedures.

Of particular interest are the so-called closure methods [31, 47, 48]. Assume one is interested in testing a given set of hypotheses $\mathcal{A}$. For each hypothesis, assign an α-level test, $\alpha \in (0, 1)$. The closure method for testing each of these hypotheses consists in rejecting hypothesis $H \in \mathcal{A}$ if, and only if,

1. H is rejected according to the α-level test.
2. All hypotheses in $\mathcal{A}$ that imply it (i.e. all $H' \subseteq H$, where $H' \in \mathcal{A}$) are rejected according to their respective α-level tests.

Besides controlling the FWER when $\mathcal{A}$ is closed under intersection [47], this method has the advantage of satisfying what is called the *coherence* property: if hypothesis H_0^1 implies hypothesis H_0^2 (i.e. $H_0^1 \subseteq H_0^2$) and the procedure rejects H_0^2, the closure method also rejects H_0^1 [12]. Although coherence is desirable, the examples we provided show that not all simultaneous procedures satisfy it.

Coherence is not the only relationship one might expect from conclusions of simultaneous hypotheses tests (to avoid confusions, from here on we call this property *monotonicity* instead, and reserve the use of the term 'coherent' for its meaning in Standard Logic, i.e. the overall logical consistency among the truth values of hypotheses that are rejected and those that are not rejected). Recently, much emphasis has been given to a different property named *consonance*, also introduced

by [12]. Informally, such a property states that when a testing procedure rejects the intersection of several hypotheses, is should also reject at least one of them marginally [39, 47, 48]. Many closure methods that respect this property have been developed [38, 54]. Finally, [25] defined a different logical property which he named *compatibility* and will be revisited later in the article. Several other consistency relationships can also be defined.

1.2 Contribution

The main goals of this work are:

1. to formalize and characterize four consistency properties for multiple test procedures;
2. to provide several procedures that satisfy them; and
3. to examine how restrictive these properties are when put together.

In Section 2, we introduce the concept of a *testing scheme*, which from now on we abbreviate to *TS*, a mathematical device that associates one test function to each hypothesis of interest. In Section 3, we formalize four consistency relations one could desire from TSs. They are *monotonicity*, *union consonance*, *intersection consonance* and *invertibility*. Next, we study some of their properties and consequences, and whether some common statistical procedures satisfy them. Finally, in Section 4, we study how restrictive these four requirements are when put together. In particular, we compare them with Lehmann's compatible schemes. Conclusions are presented in Section 5. We omit trivial demonstrations in the article.

2 Testing schemes

We start by defining a *testing scheme* (TS), a mathematical object that formalizes the notion that to each hypothesis of interest one assigns a hypothesis test (a test function). This raises the question of which are the hypotheses of interest for a given problem. This is problem dependent. However, as stated by [13], '*In some types of exploratory research it may be impossible to specify in advance the family of all potential inferences that may be of interest*.' Hence, in this work, we assume one has to assign a hypothesis test to each element of a given σ-field of the parameter space. This allows one to assign a test to each of the possible hypotheses that exist (by taking the σ-field to be the power set of the parameter space Θ), and also accommodates Bayesian procedures based on posterior probabilities, in which it is only possible to assign probabilities to some σ-fields of Θ. Also, in the case where $\theta = (\theta_0, \theta_1)$, $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$, where θ_1 are nuisance parameters [5], one can consider a σ-field of the form $\sigma(\Theta) = \sigma(\Theta_0) \times \Theta_1 = \{A \times \Theta_1 : A \in \sigma(\Theta_0)\}$. Hence, one can assign tests only to parameters of interest. This requirement is important for most results derived here. Recall that a test function is a function from the sample space $\mathcal{X}$ to $\{0, 1\}$, where 1 represents the decision of rejecting the null hypothesis and 0 represents the decision of not rejecting it. We warn the reader that this notation is standard in statistics (the most celebrated textbooks in mathematical statistics such as [6] and [43] adopt it), although it is opposite to the usual notation of the truth value of a sentence used by the logic community.

REMARK While some argue the decision 0 should be interpreted as the definitive action 'accept the hypothesis', others believe it is more appropriate to understand it as 'not reject the hypothesis', suggesting a more cautious posture over decision-making (see e.g. discussions in [20, 22, 32]). Thus, the coherence properties we define can be more or less appealing depending on which of the above

positions is adopted by a practitioner. The reader should keep these interpretations in mind when judging how reasonable each of these properties is. We return to this point later in the article.

DEFINITION 2.1 (Testing scheme; TS)
Let $\sigma(\Theta)$, the set of hypotheses to be tested, be a σ-field of the parameter space Θ. Moreover, let $\Psi = \{\phi : \mathcal{X} \to \{0,1\} : \phi \text{ is } \sigma(\mathcal{X})\text{-measurable}\}$ be the set of all test functions. A *TS* is a function $\mathcal{L} : \sigma(\Theta) \to \Psi$ that, for each hypothesis $A \in \sigma(\Theta)$, associates the test $\mathcal{L}(A) \in \Psi$ for testing A.

Hence, for hypothesis $A \in \sigma(\Theta)$ and data $x \in \mathcal{X}$, $\mathcal{L}(A)(x) = 0$ represents the decision of not rejecting A, and $\mathcal{L}(A)(x) = 1$ of rejecting it. Examples 2.2 and 2.3 illustrate this concept by introducing testing schemes induced by two traditional statistical tests. We denote the likelihood function at $\theta \in \Theta$ generated by the sample point $x \in \mathcal{X}$ by $L_x(\theta)$, which we assume to be always defined.

EXAMPLE 2.2 (Likelihood ratio tests of size α)
Let $\Theta = \mathbb{R}^d$ and $\sigma(\Theta) = \mathcal{P}(\Theta)$ be the power set of Θ. For each hypothesis $A \in \sigma(\Theta)$, let $\mathcal{L}(A) : \mathcal{X} \to \{0,1\}$ be defined by

$$\mathcal{L}(A)(x) = \mathbb{I}\left(\frac{\sup_{\theta \in A} L_x(\theta)}{\sup_{\theta \in \Theta} L_x(\theta)} \leq c_A\right),$$

where $\mathbb{I}(B)$ is the indicator function that B holds and $c_A \in [0,1]$ is chosen so that each test has size at most $\alpha \in (0,1)$ previously fixed. This is the TS that associates a likelihood ratio test of size at most α to each hypothesis $A \in \mathcal{P}(\Theta)$. ∎

EXAMPLE 2.3 (Tests based on posterior probabilities)
Assume the same set-up as Example 2.2, but now with $\sigma(\Theta)$ being the Borelians of $\mathbb{R}^d$. Assume that a prior probability $\mathbb{P}$ in $\sigma(\Theta)$ is fixed. For each $A \in \sigma(\Theta)$, let $\mathcal{L}(A) : \mathcal{X} \to \{0,1\}$ be defined by

$$\mathcal{L}(A)(x) = \mathbb{I}\left(\mathbb{P}(A|x) < \frac{1}{2}\right),$$

where $\mathbb{P}(.|x)$ is the posterior distribution of θ, given x. This is the TS that associates with each hypothesis A, the test that rejects it when its posterior probability is smaller than $1/2$. ∎

From a Bayesian decision-theoretic perspective, a hypothesis test is derived, for each sample point, by minimizing the posterior expectation of a loss function with respect to the posterior distribution of the parameters after observing the data [6]. Recall that a loss function for a test is a function $L : \{0,1\} \times \Theta \to \mathbb{R}$ that assigns to each $\theta \in \Theta$ the loss $L(d,\theta)$ for making the decision $d \in \{0,1\}$ of rejecting or not the null hypothesis. Moreover, the Bayes test is given, for each $x \in \mathcal{X}$, by $\operatorname{argmin}_{d \in \{0,1\}} \mathbb{E}[L(d,\theta)|X = x]$. Hence, in the situation of multiple tests and for a fixed probability distribution for θ, one can derive for each $A \in \sigma(\Theta)$ a Bayes test for the null hypotheses A considering a specified loss function $L_A : \{0,1\} \times \Theta \to \mathbb{R}$. This procedure is formalized by the following definition:

DEFINITION 2.4 (TS generated by a family of loss functions)
Let $(\mathcal{X} \times \Theta, \sigma(\mathcal{X} \times \Theta), \mathbb{P})$ be a Bayesian statistical model. Let $(L_A)_{A \in \sigma(\Theta)}$ be a family of loss functions, where $L_A : \{0,1\} \times \Theta \to \mathbb{R}$ is the loss function to be used to test $A \in \sigma(\Theta)$. A TS generated by the family of loss functions $(L_A)_{A \in \sigma(\Theta)}$ is any TS $\mathcal{L}$ defined over the elements of $\sigma(\Theta)$ such that, $\forall A \in \sigma(\Theta)$, $\mathcal{L}(A)$ is a Bayes test for hypothesis A against $\mathbb{P}$.

Example 2.5 illustrates this concept.

EXAMPLE 2.5 (Tests based on posterior probabilities)
Assume the same scenario as Example 2.3 and that $(L_A)_{A\in\sigma(\Theta)}$ is a family of loss functions such that $\forall A\in\sigma(\Theta)$ and $\forall\theta\in\Theta$,

$$L_A(0,\theta)=\mathbb{I}(\theta\notin A) \text{ and } L_A(1,\theta)=\mathbb{I}(\theta\in A),$$

that is, L_A is the 0-1 loss for A [42]. The testing scheme $\mathcal{L}$ defined in Example 2.3 is a TS generated by this family of loss functions. ∎

Example 2.6 shows a TS of Bayesian tests motivated by different epistemological considerations (see [51], but also [30], for a decision-theoretic motivation), the *Full Bayesian Significance Tests*, FBST [34].

EXAMPLE 2.6 (FBST testing scheme)
Let $\Theta=\mathbb{R}^d$, $\sigma(\Theta)$ be the Borelians of $\mathbb{R}^d$, and $f(\theta)$ be the prior probability density function (p.d.f.) for θ. Suppose that, for each $x\in\mathcal{X}$, there exists $f(\theta|x)$, the p.d.f. of the posterior distribution of θ, given x. For each hypothesis $A\in\sigma(\Theta)$, let

$$T_x^A=\left\{\boldsymbol{\theta}\in\Theta : f(\boldsymbol{\theta}|\boldsymbol{x})>\sup_{\theta\in A} f(\boldsymbol{\theta}|\boldsymbol{x})\right\}$$

be the set tangent to the null hypothesis and let $ev_x(A)=1-P(\theta\in T_x^A|\boldsymbol{x})$ be the Pereira–Stern evidence value for A (see [34] for a geometric motivation). One can define a TS $\mathcal{L}$ by

$$\mathcal{L}(A)(x)=\mathbb{I}(ev_x(A)\leq c),\ \forall A\in\sigma(\Theta) \text{ and } \forall x\in\mathcal{X},$$

in which $c\in[0,1]$ is fixed. In words, one does not reject the null hypothesis when its evidence is larger than c. ∎

We end this section by defining a TS generated by a point estimation procedure, an intuitive concept that plays an important role when characterizing logically coherent procedures in Section 4.

DEFINITION 2.7 (TS generated by a point estimation procedure)
Let $\widehat{\theta}:\mathcal{X}\longrightarrow\Theta$ be a point estimator. The TS generated by $\widehat{\theta}$ is defined by $\mathcal{L}(A)(x)=\mathbb{I}(\widehat{\theta}(x)\notin A)$.

Hence, we reject hypothesis A after observing x if, and only if, the point estimate for θ, $\widehat{\theta}(x)$, is not in A.

EXAMPLE 2.8 (TS generated by a point estimation procedure)
Let $\Theta=\mathbb{R}$, $\sigma(\Theta)=\mathcal{P}(\Theta)$, and assume $X_1\ldots,X_n|\theta$ c.i.i.d. $N(\theta,1)$. The TS generated by $\overline{x}$, the sample mean, rejects $A\in\sigma(\Theta)$ when $\overline{x}\notin A$. ∎

3 Consistency properties

In this section, we study four properties that one might expect from TSs to induce logically consistent tests (see Figure 1 for an illustration). Each formal definition in the sequence is preceded by an example for motivation.

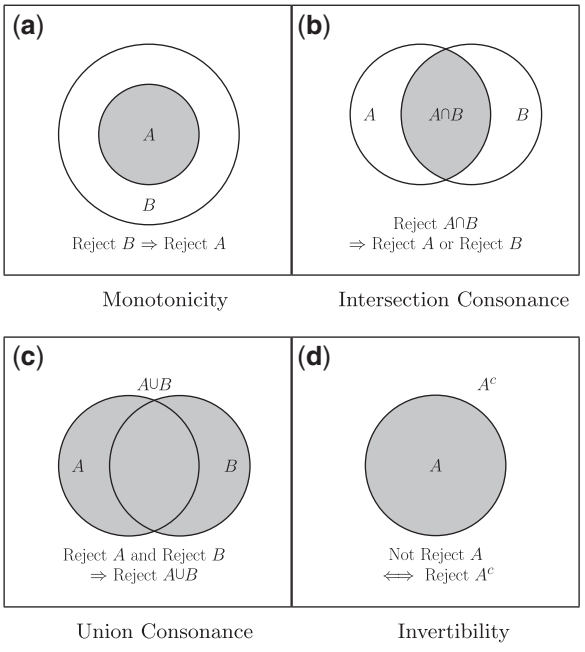

FIG. 1. Logical properties one might expect from hypotheses tests.

TABLE 1. Genotypic sample frequencies

	AA	AB	BB	Total
Case	55	83	50	188
Control	24	42	39	105

3.1 Monotonicity

The first property is related to nested hypotheses. It states that if hypothesis A *implies* hypothesis B (i.e. $A \subseteq B$), a testing scheme that rejects B should also reject A (equivalently, if it does not reject A it should not reject B either).

EXAMPLE 3.1

Suppose that in a case–control study one measures the genotype in a certain *locus* for each individual of a sample. Results are shown in Table 1. These numbers were taken from a study presented by [28] that had the aim of verifying the hypothesis that subunits of the gene $GABA_A$ contribute to a condition known as methamphetamine use disorder. Here, the set of all possible genotypes is $\{AA, AB, BB\}$. Let $\boldsymbol{\gamma} = (\gamma_{AA}, \gamma_{AB}, \gamma_{BB})$, where γ_i is the probability that an individual from the case group has genotype i. Similarly, let $\boldsymbol{\pi} = (\pi_{AA}, \pi_{AB}, \pi_{BB})$, where π_i is the probability that an individual of control group has genotype i.

In this context, two hypotheses are of interest: the hypothesis that the genotypic proportions are the same in both groups, $H_0^G : \boldsymbol{\gamma} = \boldsymbol{\pi}$, and the hypothesis that the allelic proportions are the same in both groups $H_0^A : \gamma_{AA} + \frac{1}{2}\gamma_{AB} = \pi_{AA} + \frac{1}{2}\pi_{AB}$. The p-values obtained using chi-square tests for these

hypotheses are, respectively, 0.152 and 0.069. Hence, at the level of significance $\alpha = 10\%$, the TS given by chi-square tests rejects H_0^A, but does not reject H_0^G. That is, the TS leads a practitioner to believe that the allelic proportions are different in both groups, but it does not suggest any difference between the genotypic proportions. This is absurd! If the allelic proportions are *not* the same in both groups, the genotypic proportions *cannot* be the same either. Indeed, if the latter were the same, then $\gamma_i = \pi_i$, $\forall i \in \mathbb{G}$, and hence $\theta \in H_0^A$. This example is further discussed in [18]. ∎

This example motivates the following definition, first introduced by [12] in a different setting (filter of hypotheses):

DEFINITION 3.2 (Monotonicity)
A testing scheme $\mathcal{L}$ is monotonic if

$$\forall A, B \in \sigma(\Theta),\ A \subseteq B \Rightarrow \mathcal{L}(A) \geq \mathcal{L}(B), \text{ i.e., } \forall x \in \mathcal{X},\ \mathcal{L}(A)(x) \geq \mathcal{L}(B)(x).$$

In words, if after observing x, a hypothesis is rejected by a testing scheme, any hypothesis that implies it also has to be rejected by the same scheme.

Monotonicity has received a lot of attention in the literature. It has been considered a very appealing property, whether the decision $\mathcal{L}(A)(x) = 0$ is interpreted as 'not reject the hypothesis A' or 'accept the hypothesis A'.

Theorem 3.3 shows that monotonic TSs control the FWER. Its proof is omitted as different versions were already provided in several works (e.g. [13] and [47, 48]).

THEOREM 3.3
Let $\mathcal{L}$ be a monotonic TS and assume that $\{\theta\} \in \sigma(\Theta)$, $\forall \theta \in \Theta$, i.e. the simple hypotheses are in $\sigma(\Theta)$. Then,

$$\begin{aligned} \text{FWER} &:= \sup_{\theta \in \Theta} \mathbb{P}(\text{Reject at least one correct } A \in \sigma(\Theta) | \theta) \\ &= \sup_{\theta \in \Theta} \mathbb{P}(\mathcal{L}(\{\theta\})(X) = 1 | \theta). \end{aligned}$$

Hence, if each of the tests for the simple hypotheses is of size α, FWER $\leq \alpha$, and, consequently, each hypothesis test (for a simple or composite one) will also have size α.

As discussed in the introduction, closure procedures are monotonic. Moreover, [47, 48] shows that any monotonic procedure can be constructed using the closure method. [49] showed that any non-monotonic procedure can be replaced by a monotonic one which is better in the sense that it has the same FWER as the original procedure, and rejects not only the hypotheses rejected by the first, but also potentially more of them.

Example 3.1 showed that p-values can yield non-monotonic testing schemes. The use of Bayes Factors can also result in inconsistent conclusions [24]. In fact, in Example 3.1, the Bayes Factor in favour of H_0^A is 0.28, while the Bayes Factor in favour of H_0^G is 6.63 (using independent uniform priors over the simplexes). Hence, inconsistency remains. Likelihood ratio tests with a fixed level α (Example 2.2) are also not monotonic [18]. However, the likelihood ratio statistic is. This motivates the tests proposed by [12], which we recall in the next example.

EXAMPLE 3.4 (Likelihood ratio tests with fixed threshold)
Let $c \in [0, 1]$ and define $\mathcal{L}$ by

$$\mathcal{L}(A)(x) = \mathbb{I}\left(\frac{\sup_{\theta \in A} L_x(\theta)}{\sup_{\theta \in \Theta} L_x(\theta)} \leq c \right),\ \forall A \in \sigma(\Theta) \text{ and } \forall x \in \mathcal{X}.$$

TABLE 2. Loss functions for tests of Example 3.6

	State of Nature			State of Nature	
Decision	$\theta \in H_0^A$	$\theta \notin H_0^A$	Decision	$\theta \in H_0^B$	$\theta \notin H_0^B$
0	0	1	0	0	1
1	2	0	1	1	0

This TS is monotonic. This follows from the fact that if $A, B \in \sigma(\Theta)$ are such that $A \subseteq B$ and $x \in \mathcal{X}$, then $\sup_{\theta \in A} L_x(\theta) \leq \sup_{\theta \in B} L_x(\theta)$. ∎

In this example, to attain monotonic testing schemes with likelihood ratio tests, one gives up on having common size α for each test. Some authors defend the use of the likelihood itself as a measure of evidence [3], in which cases the TS defined in Example 3.4 is appropriate.

The FBST TS defined in Example 2.6 is in some sense the Bayesian counterpart of Example 3.4 and is also monotonic.

EXAMPLE 3.5 (FBST testing scheme)
$\mathcal{L}$ defined in Example 2.6 is monotonic. In fact, let $A, B \in \sigma(\Theta)$ be such that $A \subseteq B$ and let $x \in \mathcal{X}$ be such that $\mathcal{L}(A)(x) = 0$. We have $\sup_B f(\theta|x) \geq \sup_A f(\theta|x)$. Hence, $T_x^B \subseteq T_x^A$, and, therefore, $ev_x(A) \leq ev_x(B)$, from which monotonicity holds. ∎

For a similar reason, the test developed by [33] is also monotonic.

Bayesian tests based on posterior probabilities with a fixed common cut-off (as in Example 2.3, with cut-off $1/2$), generated by a family of $0-1-c$ loss functions [43], are monotonic. This follows from monotonicity of probabilities. However, other families of loss functions may induce non-monotonic TSs of Bayesian tests: such loss functions lead to a different cut-off for each hypothesis test to be conducted. This is illustrated in Example 3.6.

EXAMPLE 3.6
Assume $X \sim Bernoulli(\theta)$, $\theta \in [0,1]$, and that we are interested in testing the following hypotheses:

$$H_0^A : \theta \leq 0.6, \text{ and } H_0^B : \theta \leq 0.7.$$

Notice that $H_0^A \subset H_0^B$. Assume we use the loss functions from Table 2.

The Bayes tests for testing H_0^A and H_0^B are, respectively,

$$\mathcal{L}(H_0^A)(x) = \mathbb{I}\left(\mathbb{P}(\theta \in H_0^A|x) \leq 1/3\right) \text{ and } \mathcal{L}(H_0^B)(x) = \mathbb{I}\left(\mathbb{P}(\theta \in H_0^B|x) \leq 1/2\right).$$

If the prior for θ is uniform and we observe $x = 1$, we have $\mathbb{P}(\theta \in H_0^A|x) = 0.36$ and $\mathbb{P}(\theta \in H_0^B|x) = 0.49$, so we do not reject H_0^A, but reject H_0^B. As $H_0^A \subseteq H_0^B$, we conclude monotonicity does not hold. Intuitively, this happens because the loss of rejecting H_0^A when $\theta \in H_0^A$ is twice as large as the loss of rejecting H_0^B when $\theta \in H_0^B$. Hence, we only reject H_0^A when there is very little evidence it holds (when compared to the evidence needed to reject H_0^B). ∎

A question then arises. What conditions must be imposed on the loss functions so that the resultant Bayesian TSs are monotonic? Next, we study monotonicity under a Bayesian Decision-Theoretic

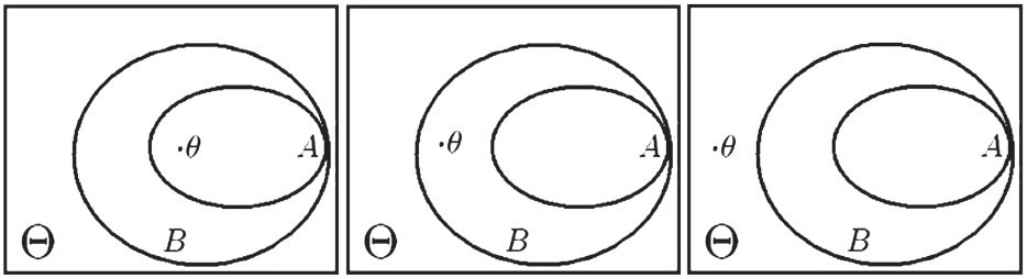

FIG. 2. Interpretation of *monotonic relative losses*: rougher errors of decisions should be assigned larger relative losses.

perspective by considering two properties for a family of loss functions $(L_A)_{A\in\sigma(\Theta)}$. We start with the following definitions:

DEFINITION 3.7 (Relative Loss)
Let L_A be a loss function for testing hypothesis A. The function $r_A:\Theta\to\mathbb{R}$ defined by $r_A(\theta)=L_A(1,\theta)-L_A(0,\theta)$ is named the *relative loss* of L_A for testing A.

The relative loss thus measures the difference between the losses of rejecting and not rejecting a given hypothesis.

DEFINITION 3.8
The family $(L_A)_{A\in\sigma(\Theta)}$ has *proper relative losses* if, for all $A\in\sigma(\Theta)$, $r_A(\theta)\geq 0$, $\forall\theta\in A$ and $r_A(\theta)\leq 0$, $\forall\theta\in A^c$.

Definition 3.8 generalizes the early definition of loss functions for hypothesis testing [43], and it states that by taking a correct decision we lose less than or the same by taking a wrong decision.

DEFINITION 3.9
The family $(L_A)_{A\in\sigma(\Theta)}$ has *monotonic relative losses* if, for all $A,B\in\sigma(\Theta)$ such that $A\subseteq B$, $r_B(\theta)\geq r_A(\theta)$, $\forall\theta\in\Theta$.

Definition 3.9 can be easily interpreted in three cases (see Figure 2). For $A,B\in\sigma(\Theta)$ with $A\subseteq B$,

- If $\theta\in A$, both A and B are true, so $(L_A)_{A\in\sigma(\Theta)}$ having monotonic relative losses reflects the situation in which the rougher error of rejecting B compared to rejecting A should be assigned a larger relative loss.
- If $\theta\in B\backslash A$, A is false and B is true. Thus, $(L_A)_{A\in\sigma(\Theta)}$ having monotonic relative losses is a natural consequence of satisfying proper relative losses.
- If $\theta\in B^c$, it can be interpreted in a similar way as the first case.

EXAMPLE 3.10
The following families of loss functions have *proper and monotonic relative losses*:

- Losses of the form of Table 3, with the restrictions that $\forall A\in\sigma(\Theta)$, $a_A=b_{A^c}$, and that $\forall A,B\in\sigma(\Theta)$ such that $A\subseteq B$, $a_A\geq a_B\geq 0$. Notice that, with these restrictions, it holds that $0\leq b_A\leq b_B$ if $A\subseteq B$.
- When Θ is equipped with a distance, say d, losses of the form $L_A(0,\theta)=f(d(\theta,A))$ and $L_A(1,\theta)=f(d(\theta,A^c))$, in which $d(\theta,A):=\inf_{a\in A}d(\theta,a)$ and f is a non-decreasing function in $\mathbb{R}_+$. ∎

TABLE 3. Example of loss function

	State of Nature	
Decision	$\theta \in A$	$\theta \in A^c$
0	0	a_A
1	b_A	0

Theorem 3.11 establishes that monotonic relative losses yield monotonic Bayesian TSs. Moreover, it shows that if all relative loss functions are proper, monotonicity of the relative losses is, in some sense, necessary for monotonicity of the resulting Bayesian TS.

THEOREM 3.11
Let $(\mathcal{X} \times \Theta, \sigma(\mathcal{X} \times \Theta), \mathbb{P})$ be a Bayesian statistical model, let $(L_A)_{A \in \sigma(\Theta)}$ be a family of loss functions, and $\mathcal{L}$ a TS generated by this family. If $\forall A \in \sigma(\Theta)$ and $\forall x \in \mathcal{X}$, $|\mathbb{E}[L_A(0,\theta)|x]| < \infty$ and $|\mathbb{E}[L_A(1,\theta)|x]| < \infty$, then:

1. If $(L_A)_{A \in \sigma(\Theta)}$ has monotonic relative losses, $\mathcal{L}$ is monotonic, whatever the prior distribution for θ is.
2. If $(L_A)_{A \in \sigma(\Theta)}$ has proper relative losses, but there exist $A, B \in \sigma(\Theta)$, with $A \subset B$, and $\theta_1 \in A$ and $\theta_2 \in B^c$ such that $r_B(\theta_i) < r_A(\theta_i)$, $i = 1, 2$, and $L_{\theta_1}(x), L_{\theta_2}(x) > 0$ $\forall x \in \mathcal{X}$, then there exists a prior distribution for which $\mathcal{L}$ is not monotonic.

See the Appendix for a proof of part 2. Monotonic relative losses are not reasonable when one prefers 'smaller' hypotheses, i.e. when the cost (relative loss) of rejecting a 'small' hypothesis is greater than that of rejecting a 'large' one, even when both are correct (as is Example 3.6). Theorem 3.11 says that this is exactly when monotonicity may not hold; otherwise monotonicity always holds. Hence, any TS derived from the loss functions of Example 3.10 is monotonic.

3.2 Intersection consonance

The second property involves testing two hypotheses A and B separately, and testing their intersection $A \cap B$. It states that if a testing scheme leads to the rejection of the intersection of these hypotheses, $A \cap B$, it should also reject *at least* one of them, A or B. The following example shows this is not always the case.

EXAMPLE 3.12 (ANOVA)
Suppose that $X_1, \ldots, X_{20}$ are i.i.d. $N(\mu_1, \sigma^2)$; $X_{21}, \ldots, X_{40}$ are i.i.d. $N(\mu_2, \sigma^2)$ and $X_{41}, \ldots, X_{60}$ are i.i.d. $N(\mu_3, \sigma^2)$. Consider the following hypotheses:

$$H_0^{(1,2,3)}: \mu_1 = \mu_2 = \mu_3 \qquad H_0^{(1,2)}: \mu_1 = \mu_2 \qquad H_0^{(1,3)}: \mu_1 = \mu_3$$

and suppose that we observe the following means and standard deviations on the data: $\overline{X}_1 = 0.15$; $S_1 = 1.09$; $\overline{X}_2 = -0.13$; $S_2 = 0.5$ $\overline{X}_3 = -0.38$; $S_3 = 0.79$. Using the likelihood ratio statistics, we have the following p-values for these hypotheses:

$$p_{H_0^{(1,2,3)}} = 0.0498 \qquad p_{H_0^{(1,2)}} = 0.2564 \qquad p_{H_0^{(1,3)}} = 0.0920.$$

Therefore, the testing scheme given by the likelihood ratio tests with common level of significance $\alpha = 5\%$ rejects $H_0^{(1,2,3)}$ but does not reject either $H_0^{(1,2)}$ or $H_0^{(1,3)}$. Hence, we conclude that the three groups do not have the same mean. However, when comparing the first with the second, the TS does not reject that they have the same mean, as well as when it compares the first with the third. It seems puzzling that the testing scheme cannot detect where the differences between the groups are. Notice that such contradiction can happen even if one makes Bonferroni corrections: The same example illustrates this if the global significance level is taken to be $\alpha = 15\%$. ∎

This contradiction is named a *consonance* contradiction by [12]. Here, we call the consistency property *intersection consonance*, as later we will introduce the *union consonance*. Several variations were defined in the literature [3, 16, 39]. Here, we present the definition of $|S|$-intersection consonance, where we use $|S|$ to denote the cardinality of set S.

DEFINITION 3.13 ($|S|$-intersection consonance)
A testing scheme $\mathcal{L}$ satisfies the $|S|$-intersection consonance if for all sets of indices I with cardinality $|I| \leq |S|$,

$$\forall \{A_i\}_{i \in I} \subseteq \sigma(\Theta) \text{ such that } \cap_{i \in I} A_i \in \sigma(\Theta), \text{ we have}$$
$$\mathcal{L}(\cap_{i \in I} A_i) \leq \max\{\mathcal{L}(A_i)\}_{i \in I}.$$

In words, if the testing scheme does not reject any of the hypotheses $\{A_i\}_{i \in I}$, it should also not reject their intersection.

In Section 4, we will specially be interested in three cases of intersection consonance, namely:

- **finite intersection consonance**. In this case, $S = \{0, 1\}$, and we only require such property to hold for a *finite* number of hypotheses. This is because taking $S = \{0, 1\}$ yields the same testing schemes as taking $S = \{0, \ldots, n\}$ for any finite natural n.
- **countable intersection consonance.** In this case, $S = \mathbb{N}$, and we only require such property to hold for a *countable* number of hypotheses.
- **complete intersection consonance**. In this case, $S = \Theta$, and we require such property to hold for any set of hypotheses with cardinality $|\Theta|$.

Although it is usually considered that consonance is not as a strong requirement as monotonicity, many consider it to be a desirable property (e.g. [4, 13, 14, 39]). As [54] points out, *'… ensuring a Multiple Test Procedure to be consonant is also important from both interpretive and mathematical statistics viewpoint'*. This is because *'the investigator is not satisfied with an overall statement, such as that there are differences "anywhere", but he wishes to determine more exactly where these differences are located'* [14].

Next, we illustrate a testing scheme that satisfies intersect consonance.

EXAMPLE 3.14
Suppose that $\{\theta\} \in \sigma(\Theta)$, $\forall \theta \in \Theta$. For each $A \in \sigma(\Theta)$, let

$$\mathcal{L}(A)(x) = \mathbb{I}\left(R(x) \nsubseteq A\right), \forall x \in \mathcal{X},$$

in which $R: \mathcal{X} \longrightarrow \mathcal{P}(\Theta)$ is a region estimator of θ. In words, the TS rejects a hypothesis if, and only if, the estimated region is not fully contained in (i.e. is not a subset of) the hypothesis of interest. $\mathcal{L}$ satisfies both the $|\Theta|$-intersection consonance and monotonicity. ∎

Many simultaneous hypotheses procedures developed satisfy intersection consonance (see e.g. [48, 49], and [38], who also discuss optimal power properties of such procedures). As noted by [12], tests that satisfy monotonicity and intersection consonance are related to union–intersection tests. More specifically, in the context of TSs, for such procedures we have that $\mathcal{L}(\cap_{i\in I}A_i)=\max_{i\in I}\mathcal{L}(A_i)$ for all I with cardinality $|I|\leq|S|$. That is, the test for $\cap_{i\in I}A_i$ is the union–intersection test based on tests for the hypotheses A_i [5], a fact that motivates several consonant schemes (e.g. [16]).

The following example shows a simple way to create TSs with intersection consonance based on tests for simple hypothesis.

EXAMPLE 3.15
For each $\theta\in\Theta$, let $\mathcal{L}(\{\theta\})$ be a hypotheses test for the simple hypothesis $\{\theta\}$. Consider the testing scheme $\mathcal{L}'$ defined by

$$\mathcal{L}'(A)(x)=1-\min_{\theta\in A^c}\mathcal{L}(\{\theta\})(x).$$

$\mathcal{L}'$ satisfies both the $|\Theta|$-intersection consonance and monotonicity. Indeed, let $A,B\in\sigma(\Theta)$, with $A\subseteq B$. As $B^c\subseteq A^c$, we have that, for every $x\in\mathcal{X}$,

$$\mathcal{L}'(A)(x):=1-\min_{\theta\in A^c}\mathcal{L}(\{\theta\})(x)\geq 1-\min_{\theta\in B^c}\mathcal{L}(\{\theta\})(x):=\mathcal{L}'(B)(x),$$

and thus monotonicity holds. Now, let $\{A_i\}_{i\in I}$, with $A_i\in\sigma(\Theta)$ be an arbitrary set of hypotheses. We have that, for every $x\in\mathcal{X}$ and for every $k\in I$

$$\begin{aligned}\mathcal{L}'(\cap_i A_i)(x):&=1-\min_{\theta\in(\cap_i A_i)^c}\mathcal{L}(\{\theta\})(x)=1-\min_{\theta\in(\cup_i A_i^c)}\mathcal{L}(\{\theta\})(x)\leq 1-\min_{\theta\in A_k^c}\mathcal{L}(\{\theta\})(x)\\&:=\mathcal{L}'(A_k)(x).\end{aligned}$$

Thus, $\mathcal{L}'(\cap_i A_i)(x)\leq\max_{i\in I}\mathcal{L}'(A_i)(x)$, from which $|\Theta|$-intersection consonance follows. Because of monotonicity, $\mathcal{L}'$ also controls the FWER [48]. ∎

Finally, if a TS satisfies both finite intersection consonance and monotonicity, it also respects the following the logical property explored by [44]: suppose in Example 3.12 we test $\mu_1=\mu_2$, $\mu_1=\mu_3$ and $\mu_2=\mu_3$. Then if a TS rejects any one of these hypotheses it also rejects at least another of them: without loss of generality, assume it rejects $\mu_1=\mu_2$. Then, by monotonicity,

$$1=\mathcal{L}(\{\mu_1=\mu_2=\mu_3\})(x)=\mathcal{L}(\{\mu_1=\mu_3\}\cap\{\mu_2=\mu_3\})(x).$$

By intersection consonance we thus have that

$$1=\mathcal{L}(\{\mu_1=\mu_3\}\cap\{\mu_2=\mu_3\})(x)=\max\{\mathcal{L}(\{\mu_1=\mu_3\})(x),\mathcal{L}(\{\mu_2=\mu_3\})(x)\},$$

which implies that the TS rejects either $\mu_1=\mu_3$ or $\mu_2=\mu_3$.

3.3 Union consonance

The third property is similar to intersection consonance; however it involves testing the union of two hypotheses. To prevent practitioners from being puzzled with results of tests, it seems advisable that if a testing scheme rejects each of the hypotheses A and B, it should also reject their union $A\cup B$. This is equivalent to stating that if it *does not* reject the union of the hypotheses, it should also retain *at least* one of them. The following example shows this is not always the case.

EXAMPLE 3.16
Suppose three candidates are running for a majority election. The proportions of electors voting for each candidate are θ_1, θ_2 and θ_3, with $\sum_{i=1}^{3}\theta_i = 1$. We are interested in testing the following four hypotheses:

$$H_0^0 : \bigcup_{i=1}^{3}\left\{\theta_i > \frac{1}{2}\right\},$$
$$H_0^1 : \left\{\theta_1 > \frac{1}{2}\right\}, \qquad H_0^2 : \left\{\theta_2 > \frac{1}{2}\right\}, \qquad H_0^3 : \left\{\theta_3 > \frac{1}{2}\right\}.$$

Hence, the null hypothesis H_0^0 is the hypothesis that one of the candidates has more than 50% of the votes, while the null hypothesis H_0^i, for $i = 1,2,3$, is the hypothesis that the i-th candidate has more than 50% of the votes. Assume we observe a sample of 410 electors. Let $X = (X_1, X_2, X_3)$, in which X_i is the number of electors in the sample that vote for candidate i, $i = 1,2,3$. Using a uniform prior for $\theta = (\theta_1, \theta_2, \theta_3)$ and assuming a multinomial distribution for $X|\theta$, if the observed sample is $x = (200, 200, 10)$, we have that $\theta|x$ is Dirichlet with parameters $(201, 201, 11)$, and therefore

$$\mathbb{P}\left(\bigcup_{i=1}^{3}\left\{\theta_i > \frac{1}{2}\right\}\middle| x\right) = 0.588;$$
$$\mathbb{P}\left(\left\{\theta_1 > \frac{1}{2}\right\}\middle| x\right) = 0.294; \quad \mathbb{P}\left(\left\{\theta_2 > \frac{1}{2}\right\}\middle| x\right) = 0.294; \quad \mathbb{P}\left(\left\{\theta_3 > \frac{1}{2}\right\}\middle| x\right) = 0.000.$$

The TS described in Example 2.3 does not reject H_0^0 but rejects H_0^i, $i = 1,2,3$. We thus have conflicting conclusions: the testing scheme leads one to conclude that one of the candidates has at least 50% of the votes (i.e., $\theta_i > 1/2$ for some i). However, separately, one concludes that each of the candidates has at most 50% of the votes (i.e., $\theta_i \leq 1/2$ for all i). ∎

We call this inconsistency lack of *union consonance*, which we formally define in what follows. To the best of the authors' knowledge, union consonance has not been formally defined in the statistics literature.

DEFINITION 3.17 ($|S|$-union consonance)
A testing scheme $\mathcal{L}$ satisfies the $|S|$-union consonance if for all sets of indices I with cardinality $|I| \leq |S|$,

$$\forall \{A_i\}_{i\in I} \subseteq \sigma(\Theta) \text{ such that } \cup_{i\in I} A_i \in \sigma(\Theta), \text{ we have } \mathcal{L}(\cup_{i\in I} A_i) \geq \min\{\mathcal{L}(A_i)\}_{i\in I}.$$

In words, if a testing scheme does not reject the union of the hypotheses $\{A_i\}_{i\in I}$, it should retain at least one of them.

Union consonance is related to the lottery paradox [23], the results of which are considered to be paradoxical by some authors, while are regarded as not contradictory by others. Thus, union consonance may not be as appealing as monotonicity at a first glance, although it has been hinted in various works. For example, the interpretation given by [11](page 1199) on the final joint decisions derived from all partial decisions implicitly suggests that union consonance is reasonable: these authors indicate one should consider $\bigcup_{A:\mathcal{L}(A)(x)=1} A$ to be the set of all parameter values rejected by the multiple procedure at hand when x is observed. Under this interpretation, it seems natural to

expect that $\mathcal{L}(\bigcup_{A:\mathcal{L}(A)(x)=1} A)(x)=1$, which is exactly what union consonance imposes. As a matter of fact, the *general partitioning principle* proposed by these authors satisfies union consonance. This is shown in the context of TSs in the first part of Theorem 3.18:

THEOREM 3.18
Assume that $\{\theta\}\in\sigma(\Theta)$, $\forall\theta\in\Theta$. Let $\mathcal{L}$ be a TS constructed as follows: for each $\theta\in\Theta$, fix a test $\mathcal{L}(\{\theta\})$. For each $A\in\sigma(\Theta)$, define

$$\mathcal{L}(A)=\min_{\theta\in A}\mathcal{L}(\{\theta\}),$$

the intersection–union test for A based on the tests for the hypotheses $\{\theta\}\subseteq A$—this is also the TS given by the general partitioning principle in [11] when the partition is the singletons of Θ, which is based on ideas by [50]. Then,

1. $\mathcal{L}$ satisfies the $|\Theta|$-union consonance as well as monotonicity.
2. Let $\mathcal{L}'$ be a TS that satisfies monotonicity, with $\mathcal{L}'(\{\theta\})=\mathcal{L}(\{\theta\})$, $\forall\theta\in\Theta$. If $\mathcal{L}'$ also satisfies $|\Theta|$-union consonance, we must have $\mathcal{L}'=\mathcal{L}$.

PROOF. Part 1. Let $\{A_i\}_{i\in I}\subseteq\sigma(\Theta)$ be such that $\cup_{i\in I}A_i\in\sigma(\Theta)$. We have that

$$\mathcal{L}(\cup_{i\in I}A_i)\overset{\text{def}}{=}\min_{\theta\in\cup_{i\in I}A_i}\mathcal{L}(\{\theta\})=\min_{i\in I}\min_{\theta\in A_i}\mathcal{L}(\{\theta\})\overset{\text{def}}{=}\min_{i\in I}\mathcal{L}(A_i).$$

Therefore, $\mathcal{L}(\cup_{i\in I}A_i)\leq\min_{i\in I}\mathcal{L}(A_i)$ and monotonicity holds. We also have that $\mathcal{L}(\cup_{i\in I}A_i)\geq\min_{i\in I}\mathcal{L}(A_i)$ and therefore, by definition, $|\Theta|$-union consonance holds.

Part 2. Let $A\in\sigma(\Theta)$. As, by hypothesis, monotonicity holds in $\mathcal{L}'$, then $\mathcal{L}'(A)\leq\min_{\theta\in A}\mathcal{L}'(\{\theta\})$. Analogously, by $|\Theta|$-union consonance, $\mathcal{L}'(A)\geq\min_{\theta\in A}\mathcal{L}'(\{\theta\})$. Therefore, $\mathcal{L}'(A)=\min_{\theta\in A}\mathcal{L}'(\{\theta\})=\min_{\theta\in A}\mathcal{L}(\{\theta\})=\mathcal{L}(A)$, $\forall A\in\sigma(\Theta)$. ∎

Theorem 3.18 shows that to create a TS that satisfies monotonicity and union consonance simultaneously, it is only necessary to define the tests for the simple hypotheses and consider intersection–union tests derived from them. The second part of Theorem 3.18 asserts that such testing scheme is the unique extension of the above-mentioned tests assigned to simple hypotheses to a TS that is monotonic and satisfies union consonance. In other words, when a TS satisfies arbitrary union consonance and monotonicity, its behaviour is completely determined by its behaviour on the singletons. Moreover, testing schemes created according to Theorem 3.18 control the FWER. This follows from Theorem 3.3.

EXAMPLE 3.19
For each $A\in\sigma(\Theta)$, let

$$\mathcal{L}(A)(x)=\mathbb{I}\left(R(x)\bigcap A=\emptyset\right),\ \forall x\in\mathcal{X},$$

in which $R:\mathcal{X}\longrightarrow\mathcal{P}(\Theta)$ is a region estimator of θ. In words, the TS $\mathcal{L}$ rejects a hypothesis if, and only if, the estimated region does not intersect the hypothesis of interest. This very intuitive procedure was proposed by [1] focusing on classical confidence regions. It is straightforward to show $\mathcal{L}$ satisfies $|\Theta|$-union consonance. Also, [13] noticed it is monotonic and hence controls FWER (Theorem 3.18). In particular, if $\{\theta\}\in\sigma(\Theta)$, $\forall\theta\in\Theta$, and $R(X)$ has confidence $1-\alpha$ (i.e., $\mathbb{P}(\theta\in R(X)|\theta)\geq 1-\alpha$, $\forall\theta\in\Theta$), the tests for each of the simple hypotheses $\{\theta\}$ have level α. ∎

Notice that the TS in Example 3.19 is composed of intersection–union tests based on tests of the form $\mathcal{L}(\{\theta\})(x)=\mathbb{I}(\theta\notin R(x))$, $\theta\in\Theta$, as in Theorem 3.18. The next theorem shows that all TSs that satisfy both monotonicity and complete intersection consonance must be of the form defined in Example 3.19.

THEOREM 3.20
Assume that $\{\theta\}\in\sigma(\Theta)$, $\forall\theta\in\Theta$. Then any TS defined on $\sigma(\Theta)$ satisfies both monotonicity and complete intersection consonance if, and only if, there exists a region estimator $R:\mathcal{X}\longrightarrow\mathcal{P}(\Theta)$ such that

$$\mathcal{L}(A)(x)=\mathbb{I}\left(R(x)\bigcap A=\emptyset\right),\ \forall x\in\mathcal{X}.$$

PROOF. Because the 'if' direction is trivial, we only prove the 'only if' direction.

Let $\mathcal{L}$ be a TS and for $x\in\mathcal{X}$ let $R(x)=\{\theta\in\Theta:\mathcal{L}_{\{\theta\}}(x)=0\}$. We have to show that, for each $A\in\sigma(\Theta)$,$\mathcal{L}(A)(x)=\mathbb{I}(R(x)\bigcap A=\emptyset)$. It is trivial to show this is true for hypotheses A that are singletons. Using this fact and that, because $\mathcal{L}$ satisfies union consonance and monotonicity, $\mathcal{L}(\cup_{\theta\in A}\{\theta\})(x)=\min_{\theta\in A}\mathcal{L}(\{\theta\})(x)$, we have that

$$\begin{aligned}\mathcal{L}(A)(x)&=\mathcal{L}(\cup_{\theta\in A}\{\theta\})(x)=\min_{\theta\in A}\mathcal{L}(\{\theta\})(x)=\min_{\theta\in A}\mathbb{I}\left(R(x)\bigcap\{\theta\}=\emptyset\right)\\&=\mathbb{I}\left(R(x)\bigcap A=\emptyset\right).\end{aligned}$$

In practice, procedures that satisfy both union consonance and monotonicity are usually easier to implement than the traditional closure method described in the introduction. This is because only tests for the simple hypotheses have to be constructed. If Θ is finite, it requires only $|\Theta|$ operations (instead of $2^{|\Theta|}$, as in the case of the closure method when all tests result in rejections). Such procedures are also easy to implement when Θ is continuous if confidence regions can be easily built, as in the following example:

EXAMPLE 3.21 (ANOVA)
Suppose that $X_{k,1},\ldots,X_{k,n_k}$ are i.i.d. $N(\mu_k,\sigma^2)$, $k=1,\ldots,g$, conditionally on $\mu_1,\ldots,\mu_g,\sigma^2$, and that $X_{i,j}$ is independent of $X_{k,l}$ $\forall i\neq k$ and $\forall j,l$. Here $X_{i,j}$ represents the measurement made on the j-th sample unit of the $i-$th group. A confidence region for $(\mu_1,\ldots,\mu_g)$ of confidence at least $1-\alpha$ presented by [21] associates to the sample point x the region $R(x)$ given by

$$\left\{(\mu_1,\ldots,\mu_g)\in\mathbb{R}^g:\forall k\neq l\right.$$
$$\left.\mu_k-\mu_l\in\left[\bar{x}_k-\bar{x}_l\ \pm t_{n-g}\left(\frac{\alpha}{g(g-1)}\right)\sqrt{\frac{s^2}{n-g}\left(\frac{1}{n_k}+\frac{1}{n_l}\right)}\right],1\leq k,l\leq g\right\},$$

where $n=n_1+\ldots+n_g$, $\bar{x}_k$ is the sample average of the k-th group, $s^2=(n_1-1)s_1^2+\ldots+(n_g-1)s_g^2$, where s_k^2 is the sample variance of k-th group, and $t_d(\alpha)$ denotes the α percentile of a t distribution with d degrees of freedom. Plugging the region estimator R above in the TS defined in Example 3.19 yields a TS that is monotonic, satisfies $|\Theta|$-union consonance, and controls the FWER. In this way, it is possible to test all hypotheses of interest in an ANVA problem while preserving these properties. Notice that we are treating σ^2 as a nuisance parameter (see Section 2). ∎

If one is not interested in controlling the size of the tests, other procedures can be built. Two examples are shown below.

EXAMPLE 3.22 (Likelihood ratio tests with fixed threshold)
The TS of Example 3.4 was already shown to satisfy monotonicity. If $\{\theta\} \in \sigma(\Theta)$, $\forall \theta \in \Theta$, it also satisfies $|\Theta|$-union consonance. In fact, let $A \in \sigma(\Theta)$. We have

$$\mathcal{L}(A)(x) = \mathbb{I}\left(\frac{\sup_{\theta \in A} L_x(\theta)}{\sup_{\theta \in \Theta} L_x(\theta)} \leq c\right) = \min_{\theta_0 \in A} \mathbb{I}\left(\frac{L_x(\theta_0)}{\sup_{\theta \in \Theta} L_x(\theta)} \leq c\right) \stackrel{\text{def}}{=} \min_{\theta_0 \in A} \mathcal{L}(\{\theta_0\})(x).$$

The result follows from the first part of Theorem 3.18. ∎

There are also Bayesian tests that are in accordance with union consonance. Although testing schemes based on posterior probabilities with a fixed threshold (Example 3.16) do not respect union consonance, in general FBST testing schemes do satisfy it:

EXAMPLE 3.23 (FBST testing schemes)
Example 2.6 shows that a FBST TS satisfies monotonicity. It can also be shown that it satisfies $|\Theta|$-union consonance, provided that $\forall x \in \chi$ and $\forall a \in \mathbb{R}^+$, $\mathbb{P}(\{\theta : f(\theta|x) = a\}|x) = 0$ [17]. This TS is a particular case of the TSs described in Example 3.19: it can be shown that this TS is equivalent to

$$\mathcal{L}(A)(x) = \mathbb{I}\left(A \bigcap \text{HPD}_c^x = \emptyset\right),$$

where HPD_c^x is the Highest Posterior Probability Density region [19] with probability $1-c$, based on observation x. Hence, the FBST procedure can be efficiently implemented by constructing the posterior $(1-c)-$HPD for θ and not rejecting all hypotheses that intersect it. In a sense, an FBST TS extends Lindley's tests for simple hypotheses [29], according to intersection–union procedures in Theorem 3.18. ∎

3.4 Invertibility

The following example is traditional in introductory statistics courses and illustrates the difference that exists between choosing the labels 'null hypothesis' and 'alternative hypothesis' under the classical approach to inference.

EXAMPLE 3.24
Suppose that $X|\theta \sim Normal(\theta, 1)$ and that one wants to test the following null hypotheses:

$$H_0^{\leq}: \theta \leq 0$$
$$H_0^{>}: \theta > 0$$

The Uniformly Most Powerful (UMP) Tests for these hypotheses have the following critical regions, at the level 5%, respectively:

$$\{x \in \mathbb{R} : x > 1.64\} \text{ and } \{x \in \mathbb{R} : x < -1.64\}.$$

Hence, if we observe $x = 1.0$, a TS that comprises these UMP tests does not reject either that the mean is less than or equal to 0 ($H_0^{\leq}$) or that it is greater than 0 ($H_0^{>}$). That is, on one hand, $x = 1.0$ does not bring enough evidence in favour of $\mathbb{R}_+^*$; on the other hand, it suggests $\mathbb{R}_+^*$ cannot be rejected. Therefore, the conclusion drawn from the sample observation about a hypothesis of interest strongly depends on whether it is considered as the null or the alternative hypothesis. ∎

The next definition formalizes the notion of simultaneous tests independent of the labels 'null' and 'alternative' for the hypotheses of interest.

DEFINITION 3.25 (Invertibility)
A testing scheme $\mathcal{L}$ satisfies invertibility if

$$\forall A \in \sigma(\Theta),\ \mathcal{L}(A) = 1 - \mathcal{L}(A^c).$$

In words, it is irrelevant which hypothesis is labelled as null and which is labelled as alternative.

While invertibility is typically considered to be reasonable from a Bayesian decision-theoretic standpoint (e.g. [37], Section 5.3; [43], Section 4.1.1.), it is usually not attractive for most advocates of the frequentist theory due to the interpretation of the decision $\mathcal{L}(A)(x) = 0$ as a 'not reject' rather than an 'accept'. We note, however, that there exist examples where a TS *rejects* both the null and the alternative hypotheses:

EXAMPLE 3.26
Suppose that $X|\theta \sim Normal(\theta, 1)$, and consider the parameter space $\Theta = \{-3, 3\}$. Assume one wants to test the following null hypotheses:

$$H_0^A\colon \theta = 3 \quad \text{and} \quad H_0^B\colon \theta = -3$$

The Neyman–Pearson most powerful tests for these hypotheses have the following critical regions, at the level 5%, respectively:

$$\{x \in \mathbb{R} : x < 1.35\} \quad \text{and} \quad \{x \in \mathbb{R} : x > -1.35\}.$$

Hence, if we observe $x = -0.5$, the testing scheme rejects both H_0^A and H_0^B, even though $H_0^A \cup H_0^B = \Theta$. Considering the interpretation of results of simultaneous test procedures from [11], this would lead one to decide that

$$\theta \in \left(\Theta \setminus H_0^A\right) \bigcap \left(\Theta \setminus H_0^B\right) = \emptyset.$$

■

The last example illustrates tests that are against Lehmann's principle of *compatibility of the first kind* [26, 48], which states that, for every x, the intersection of the complements of the rejected hypotheses should not be empty, i.e.

$$\bigcap_{A \in \sigma(\Theta):\mathcal{L}(A)(x)=1} A^c \neq \emptyset.$$

Not surprisingly, we will also see in the next section that in the TS framework, invertibility is implied by intersection and union consonances.

The next example illustrates a TS that respects invertibility.

EXAMPLE 3.27
Suppose that $(L_A)_{A \in \sigma(\Theta)}$ is a family of loss functions with

$$L_A(0, \theta) = a_A \mathbb{I}(\theta \notin A) \text{ and } L_A(1, \theta) = b_A \mathbb{I}(\theta \in A),\ \forall \theta \in \Theta,$$

with $a_A = b_{A^c} > 0$, $\forall A \in \sigma(\Theta)$. Let $\theta_0 = \theta_0(x) \in \Theta$ and $\mathcal{L}$ be defined as

$$\mathcal{L}(A)(x) = \mathbb{I}\left(\mathbb{P}(A|x) < \frac{a_A}{a_A + b_A}\right) + \mathbb{I}\left(\mathbb{P}(A|x) = \frac{a_A}{a_A + b_A} \text{ and } \theta_0 \notin A\right),$$

$\forall A \in \sigma(\Theta)$ and $\forall x \in \mathcal{X}$. In words, we reject A whenever its posterior probability is smaller than $a_A/(a_A + b_A)$, or its posterior probability is $a_A/(a_A + b_A)$ and θ_0 (which may depend on x) is not in A. $\mathcal{L}$ is a Bayesian TS generated by the family $(L_A)_{A \in \sigma(\Theta)}$. This TS satisfies both invertibility and monotonicity. Notice that when $\mathbb{P}(A|x) = a_A/(a_A + b_A)$, the decision to not reject A has the same expected loss as the decision of rejecting A. This TS was chosen because among all testing schemes generated by $(L_A)_{A \in \sigma(\Theta)}$, which are equivalent from a decision-theoretic point of view, it satisfies invertibility. Of course, other TSs derived from $(L_A)_{A \in \sigma(\Theta)}$ do as well. ∎

Example 3.27 can be generalized. In fact, one can verify that any family of loss functions $(L_A)_{A \in \sigma(\Theta)}$ that satisfies $L_A(0, \theta) = L_{A^c}(1, \theta)$, $\forall A \in \sigma(\Theta)$ and $\forall \theta \in \Theta$, generates TSs that respect invertibility [46]. This restriction on the loss functions implies that a type I error for testing A has to be penalized in the same way as a type II error for testing A^c.

Example 3.28
Any TS generated by a point estimation procedure (recall Definition 2.7 and Example 2.8) is invertible. Moreover, such TSs also satisfy monotonicity, $|\Theta|$-intersection and $|\Theta|$-union consonances. ∎

4 How restrictive are the consistency properties?

In Section 3, we studied four logical properties one may expect for testing schemes. We also provided results and examples with useful schemes that respect two of these conditions simultaneously (e.g Theorem 3.18, Examples 3.15, 3.19, 3.22, 3.23 and 3.27). Here, we show that requiring more than two of such properties to hold simultaneously is very restrictive: under quite general conditions, TSs that satisfy them are always generated by point estimation procedures.

We start by recalling the concept of compatibility of a multiple test procedure, introduced by [25] and here adapted to TSs.

Definition 4.1 (Compatible TS)
A TS $\mathcal{L}$ is compatible if $\forall x \in \mathcal{X}$

$$\bigcap_{A \in \sigma(\Theta)} A^{\mathcal{L}(A)(x)} \neq \emptyset,$$

where $A^0 \stackrel{\text{def}}{=} A$ and $A^1 \stackrel{\text{def}}{=} A^c$, for $A \in \sigma(\Theta)$.

In words, a testing scheme is compatible when no incoherences are allowed: the intersection of the accepted sets (hypotheses such that $\mathcal{L}(A)(x) = 0$) with the complements of the rejected ones (hypotheses such that $\mathcal{L}(A)(x) = 1$) cannot be empty. Compatibility has been considered too strong by many authors [48], including Lehmann himself [26], who provides the less stringent definition of *compatibility of the first kind* (recall Section 3.4) motivated by the fact that one might interpret the result of a test $\mathcal{L}(A)(x) = 0$ as 'not reject' rather than 'accept'. In fact, when $\{\theta\} \in \sigma(\Theta)$, $\forall \theta \in \Theta$, it is straightforward to show that $\mathcal{L}$ is compatible if, and only if, $\mathcal{L}$ is generated by a point estimation procedure (recall Definition 2.7).

We will now put together the properties presented in Section 3 with the goal of understanding how restrictive such requirements are when compared to those of a compatible TS. We begin with the following definition:

DEFINITION 4.2 (TS of type $|S|$)
We say a TS is of type $|S|$ if it satisfies the four properties from Section 3: monotonicity, $|S|$-intersection consonance, $|S|$-union consonance and invertibility.

The following theorem shows alternative characterizations of such testing schemes.

THEOREM 4.3
Let S be $\{0,1\}, \mathbb{N}$, or Θ. The following are equivalent:

1. $\mathcal{L}$ is of type $|S|$;
2. $\mathcal{L}$ satisfies monotonicity, $|S|$-intersection consonance and invertibility;
3. $\mathcal{L}$ satisfies monotonicity, $|S|$-union consonance and invertibility; and
4. $\mathcal{L}(\emptyset)=1$, $\mathcal{L}(\Theta)=0$, and $\mathcal{L}$ satisfies $|S|$-intersection and $|S|$-union consonance;

For the case $S=\{0,1\}$, we also have the additional equivalence:

5. $\forall\{A_1,\ldots,A_n\}$ finite measurable partition of Θ,

$$\sum_{i=1}^{n}(1-\mathcal{L}(A_i))=1.$$

That is, one, and only one, A_i is not rejected by the TS. A similar equivalence holds when $S=\mathbb{N}$ and Θ is partitioned into a countable number of sets.

PROOF. The implications '*1* $\Rightarrow$ *2*', '*1* $\Rightarrow$ *3*' and '*1* $\Rightarrow$ *4*' are trivial to show. We thus start by proving that '*2* $\Rightarrow$ *1*'. We just have to show that *2* $\Rightarrow$ $|S|$-union consonance. Let I be a set of indices such that $|I|\leq|S|$. Let $\{A_i\}_{i\in I}\subseteq\sigma(\Theta)$ be such that $\cup_i A_i\in\sigma(\Theta)$. By $|S|$-intersection consonance, monotonicity and invertibility, we have that

$$\mathcal{L}(\cup_i A_i)=1-\mathcal{L}(\cap A_i^c)=1-\max_i\mathcal{L}(A_i^c)=1-\max_i(1-\mathcal{L}(A_i))=\min_i\mathcal{L}(A_i),$$

which implies $|S|$-union consonance holds. A similar proof shows that '*3* $\Rightarrow$ *1*'.

We now show that '*4* $\Rightarrow$ *1*.' To verify that invertibility holds, let $A\in\sigma(\Theta)$. We have, by intersection consonance and by $\mathcal{L}(\emptyset)=1$, that $1-(1-\mathcal{L}(A))(1-\mathcal{L}(A^c))=\mathcal{L}(A\cap A^c)=\mathcal{L}(\emptyset)=1$. Hence, $(1-\mathcal{L}(A))(1-\mathcal{L}(A^c))=0$, so that $\mathcal{L}(A)\mathcal{L}(A^c)=\mathcal{L}(A)+\mathcal{L}(A^c)-1$. But, by union consonance and by $\mathcal{L}(\Theta)=0$, $\mathcal{L}(A)\mathcal{L}(A^c)=\mathcal{L}(A\cup A^c)=\mathcal{L}(\Theta)=0$. Therefore, $\mathcal{L}(A)+\mathcal{L}(A^c)-1=0$ and invertibility holds.

To verify monotonicity, let $A,B\in\sigma(\Theta)$, with $A\subseteq B$, and $x\in\chi$. If $\mathcal{L}(A)(x)=0$, by invertibility, $\mathcal{L}(A^c)(x)=1-\mathcal{L}(A)(x)=1$. As $\mathcal{L}(A\cap B^c)(x)=\mathcal{L}(\emptyset)(x)=1$ and because intersection consonance holds, $1=\mathcal{L}(A\cap B^c)(x)\leq 1-(1-\mathcal{L}(A)(x))(1-\mathcal{L}(B^c)(x))=1-1(1-\mathcal{L}(B^c)(x))$, so that $\mathcal{L}(B^c)(x)=1$. Using invertibility again, $\mathcal{L}(B)(x)=0$, and hence monotonicity holds.

We now show '*5* $\Rightarrow$ *3*' for the $|\{0,1\}|-$union consonance case. A similar proof works for the $|\mathbb{N}|-$union consonance case. Let $A\in\sigma(\Theta)$. Consider the partition $A_1=A$ e $A_2=A^c$. We have $(1-\mathcal{L}(A))+(1-\mathcal{L}(A^c))=1$, so that $\mathcal{L}(A)=1-\mathcal{L}(A^c)$ and, therefore, invertibility holds.

Now, let $A,B\in\sigma(\Theta)$, with $A\subseteq B$, and $x\in\chi$. Without loss of generality, consider $\mathcal{L}(A)(x)=0$. Let us consider the partition $A_1'=A$, $A_2'=B\backslash A$ and $A_3'=(A\cup B)^c$. By hypothesis, we have

$(1-\mathcal{L}(A)(x))+(1-\mathcal{L}(B\backslash A))(x)+(1-\mathcal{L}((A\cup B)^c)(x))=1$. Since $\mathcal{L}(A)(x)=0$, it follows that $\mathcal{L}((A\cup B)^c)(x)=1$, and, by invertibility, $\mathcal{L}(A\cup B)(x)=\mathcal{L}(B)(x)=0$. Therefore, monotonicity holds.

Considering again the partition A'_1,A'_2,A'_3, but assuming that $\mathcal{L}(A\cup B)(x)=0$, by invertibility we have that $\mathcal{L}((A\cup B)^c)(x)=1$. Hence, by hypothesis, or $\mathcal{L}(A)(x)=0$ or $\mathcal{L}(B\backslash A)(x)=0$. In the second case, by monotonicity, $\mathcal{L}(B)(x)=0$. Therefore, $\mathcal{L}(A\cup B)(x)=0\Rightarrow\mathcal{L}(A)(x)\mathcal{L}(B)(x)=0$, and finite union consonance holds.

We now show that '*1*$\Rightarrow$*5*' and hence finish the proof of the theorem. We only show it for the case $S=\{0,1\}$, but the case $S=\mathbb{N}$ has a similar proof. Let $\{A_1,\ldots,A_n\}$ be a finite measurable partition of Θ, and $x\in\chi$. We have, by finite union consonance and equivalence (*4*) from this theorem, that $\min_i\mathcal{L}(A_i)\leq\mathcal{L}\left(\bigcup_{i=1}^n A_i\right)=\mathcal{L}(\Theta)=0$, so that $\exists i_0\in\{1,\ldots,n\}$ such that $\mathcal{L}(A_{i_0})(x)=0$. But since $A_j\cap A_{i_0}=\emptyset$ for $j\neq i_0$, we have $\mathcal{L}(A_j)(x)=1$ for all $j\neq i_0$, by finite intersection consonance and equivalence (4). Therefore, $\sum_{i=1}^n(1-\mathcal{L}(A_i)(x))=1$, concluding the proof.

Of particular interest is characterization (*4*), which does not involve invertibility, controversial among advocates of frequentist methods (as a matter of fact, invertibility is (nearly) a consequence of union and intersection consonances due to Theorem 4.3). Moreover, characterizations (*2*) and (*3*) show that under invertibility and monotonicity, requiring union consonance is equivalent to requiring intersection consonance.

The following theorem shows that the concept of a TS of type $|\Theta|$ is as strong as that of a compatible TS.

THEOREM 4.4
If $\{\theta\}\in\sigma(\Theta)$, $\forall\theta\in\Theta$, any TS defined over $\sigma(\Theta)$ is of type $|\Theta|$ if, and only if, it is compatible.

As the proof of this fact is not difficult, we omit it for the sake of brevity.

It does follows that *the only examples of TSs of type* $|\Theta|$ *are those generated by point estimation procedures*:

COROLLARY 4.5
If $\{\theta\}\in\sigma(\Theta)$, $\forall\theta\in\Theta$, any TS defined over $\sigma(\Theta)$ is of type $|\Theta|$ if, and only if, it is generated by a point estimation procedure.

In the remainder of the section, we investigate whether this is also true when $S=\mathbb{N}$ or $S=\{0,1\}$.

The following theorem shows that, under some conditions, the only TSs of type $|\mathbb{N}|$ are *also* the ones generated by a point estimation procedure. Hence, under these conditions, TSs of type $|\mathbb{N}|$ are the same as compatible TSs, which are, as we argued, the same as TSs of type $|\Theta|$.

THEOREM 4.6
Assume there exists a separable topology $\tau\subseteq\sigma(\Theta)$ over Θ. Then $\mathcal{L}$ is of type $|\mathbb{N}|$ if, and only if, it is generated by a point estimation procedure.

PROOF. We prove the necessary condition only (the reverse is immediate). Let $x\in\mathcal{X}$. For each $n\in\mathbb{N}$, let $\mathcal{A}_n=\{B(\theta,\frac{1}{n}):\theta\in\Theta\}$, where $B(\theta,\frac{1}{n})$ is a ball with radius $\frac{1}{n}$ and centre θ (and distance induced by τ; see Figure 3 for an illustration). As each $\mathcal{A}_n$ covers Θ and τ is separable, for each n there exists a countable subset $\mathcal{A}_n^*\subseteq\mathcal{A}_n$ that covers Θ. Moreover, by countable union consonance, $\Pi_{D\in\mathcal{A}_n^*}\mathcal{L}(D)(x)=\mathcal{L}(\cup_{D\in\mathcal{A}_n^*}D)(x)=\mathcal{L}(\Theta)(x)=0$ (Theorem 4.3 part *4.*), so that for each n there exists a ball $D_n\in\mathcal{A}_n^*$ such that $\mathcal{L}(D_n)(x)=0$. By countable intersection consonance, $\mathcal{L}(\bigcap_n D_n)(x)=0$. Theorem 4.3 part *4.* implies $\bigcap_n D_n\neq\emptyset$. The proof is concluded by noticing that as the radius goes to 0 as $n\longrightarrow\infty$, $\bigcap_n D_n$ is a unitary set and that different x's can generate different D_n's. ∎

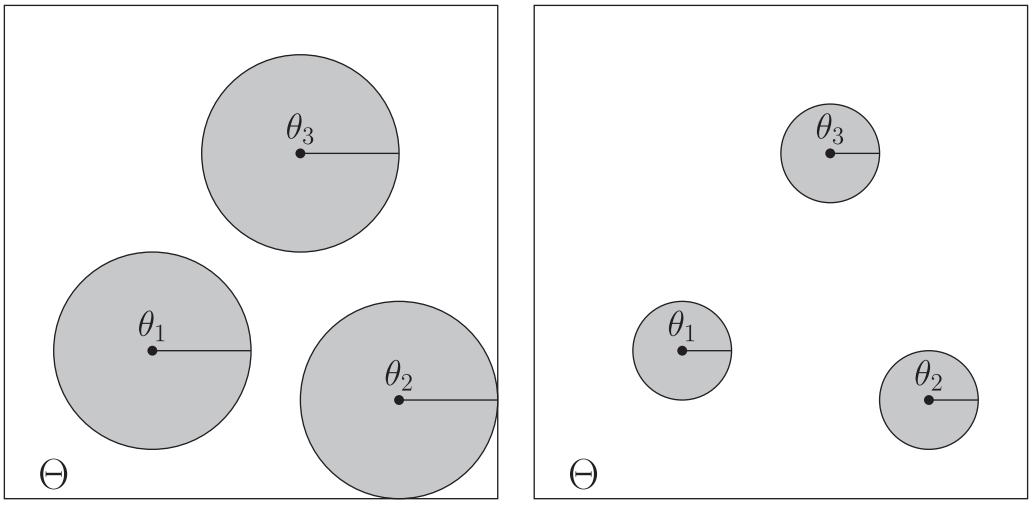

FIG. 3. Examples of sets in $\mathcal{A}_n$ from the proof of Theorem 4.6 for two different n's.

COROLLARY 4.7
If $\Theta = \mathbb{R}^d$, a TS $\mathcal{L}$ defined over any sigma-field $\sigma(\Theta)$ that contains the Borel sets is of type $|\mathbb{N}|$ if, and only if, it is generated by a point estimation procedure.

Hence, under some conditions on Θ and $\sigma(\Theta)$, we have that compatible TSs, TSs of type $|\Theta|$, TSs of type $|\mathbb{N}|$ and TSs generated by a point estimation procedure are equivalent. Theorem 4.6 and Corollary 4.7 also formally link, in a sense, point estimation and hypothesis testing. In the vast statistical literature, these two celebrated problems are most of the times treated separately (this is not the case of region estimation and hypothesis testing, as discussed in Section 3; see also [50]). Theorem 4.6 asserts that a practitioner that desires to use testing schemes of type $|\mathbb{N}|$ cannot decide, e.g. that an unknown proportion of interest is at most 50% and estimate it as 52% on the basis of the same sample information.

Are TSs of type $|\mathbb{N}|$ in fact more restrictive than TSs of type $|\{0,1\}|$? The following theorem shows that the answer is yes.

THEOREM 4.8
Assume that $\Theta = \mathbb{R}^d$ and that $\sigma(\Theta)$ contains the Borelians of Θ. There exists a TS of type $|\{0,1\}|$ which is not of type $|\mathbb{N}|$. In particular, if $\sigma(\Theta) = \mathcal{P}(\Theta)$, this existence is equivalent to the existence of a non-trivial ultrafilter over Θ.

The proof of Theorem 4.8 relies on the fact that, for a fixed $x \in \mathcal{X}$, the set $\mathbb{F}_x = \{A \in \sigma(\Theta) : \mathcal{L}(A)(x) = 0\}$ is an ultrafilter over Θ if, and only if, it is of type $|\{0,1\}|$. Moreover, it is a *trivial* ultrafilter if, and only if, it is generated by a point estimation procedure. The theorem follows from the existence of non-trivial ultrafilters [8]. It is not possible, however, to prove the existence of a non-trivial ultrafilter using only the Zermelo–Fraenkel axioms. One needs more axioms such as e.g. the Axiom of Choice [8]. Hence, it is not possible to construct 'explicit examples' of such testing schemes (see e.g. [41]). Therefore, when $\sigma(\Theta) = \mathcal{P}(\Theta)$, essentially all TSs of type $|\{0,1\}|$ that can be built are TSs generated by point estimation procedures. It is still an open question whether this is true when $\sigma(\Theta) \subsetneq \mathcal{P}(\Theta)$.

In summary, under the conditions of the theorems stated, compatible TSs are equivalent to TSs of type $|\Theta|$, TSs of type $|\mathbb{N}|$ and point estimation procedures. Moreover, although they are not

equivalent to TSs of type $|\{0,1\}|$, constructing explicit examples of TSs of type $|\{0,1\}|$ that are not compatible is not possible and, therefore, in practice they are also equivalent.

5 Discussion and conclusions

We introduced the concept of a testing scheme. Such a concept allows one to define several coherence properties that might be expected from simultaneous hypotheses tests. In particular, we studied four properties: monotonicity (also known as *coherence*, [12]), intersection consonance, union consonance and invertibility. Among these, monotonicity is the one that has been most emphasized in the literature (in particular, due to closure procedures), followed by intersection consonance. Union consonance has already been suggested, although not adequately formalized. We showed necessary and sufficient conditions for a testing scheme to be monotonic from a Bayesian decision-theoretic perspective. We also gave examples of testing schemes that satisfy each of the properties that were defined. Moreover, we gave general procedures that allow one to build schemes that satisfy monotonicity and consonance (both for union and intersection) simultaneously. Finally, we showed that when put together, these properties are very restrictive: testing schemes that satisfy (any) three of these properties are essentially equivalent to schemes generated by point estimation procedures. This is also essentially the same when both consonances are required.

The fact that the consistency properties are too restrictive when put together suggests that a practitioner may abandon two or more of these properties when performing simultaneous tests procedures, and then choose a testing scheme that combines attainment of some optimality criteria (e.g. controlling the FWER or requiring the TS to be a Bayesian TS derived from an adequate family of loss functions) with agreement to the logical consistency properties he finds more important. We provided several examples that illustrate how this can be done. In particular, the necessary and sufficient conditions for monotonicity under a decision-theoretic perspective we provided shed some light on when such property may be considered to be reasonable for a given problem. Alternatively, a practitioner might want to use a testing scheme based on a sensible point estimation procedure if monotonicity, invertibility and consonance are all of primary importance.

From another angle, the incompatibility between full logical consistency and the achievement of statistical optimality may lead one to question whether the performance of simultaneous test procedures is the most adequate way to report conclusions about a parameter from data. For instance, under the Bayesian viewpoint, many believe the most complete inference one can make about a parameter is its full posterior distribution, considered to be more informative than the list of all hypotheses that should be rejected according to decision-theoretic criteria.

Several problems are open. From a Bayesian decision-theoretic perspective, an alternative way to proceed when dealing with several hypotheses tests is to consider a single decision problem with decision space $\{0,1\}^{\sigma(\Theta)}$ taking into account joint loss functions rather than TSs. This is done by e.g. [24] and [7] for a finite number of hypotheses. Which constraints are necessary on such loss functions so that logical properties of interest are preserved?

A different approach that can be taken is that instead of considering decisions in the space $\{0,1\}$, one can create rules taking values on a decision space with three elements: accept a hypothesis of interest, reject it or do not accept or reject it, the so-called 'agnostic' tests. See e.g. [36]. One can then ask which coherence properties are expected in this framework, which is similar to the one presented by [27]. This approach also seems to be interesting as it naturally deals with the question of how (and to what extent) 'not rejecting H' is different from 'accepting H', maybe allowing a broader consensus on properties expected for simultaneous test procedures by different practitioners.

Funding

This work was supported by *Fundação de Amparo à Pesquisa do Estado de São Paulo* [2009/03385-5] and *Conselho Nacional de Pesquisa e Desenvolvimento Científico e Tecnológico* [131982/2009-5, 200959/2010-7].

Acknowledgements

The authors are thankful for Victor Fossaluza, Verónica Andrea González-López, Jay B. Kadane, Tiago Mendonça, Carlos Alberto de Bragança Pereira, Teddy Seidenfeld, Gustavo Miranda da Silva, Julio Michael Stern, Lea Veras and Sergio Wechsler for their interesting comments and suggestions on this paper. They are especially grateful for Rafael Bassi Stern for the discussions and for helping in the generalization that resulted in Theorem 4.6, originally Corollary 4.7, and to the two referees for the insightful comments that helped improve the quality of the paper.

References

[1] J. Aitchison. Confidence-region tests. *Journal of the Royal Statistical Society. Series B*, **26**, 462–476, 1964.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300, 1995.

[3] D. R. Bickel. The strength of statistical evidence for composite hypotheses with an application to multiple comparisons. *COBRA Preprint Series*, **22**, 1–37, 2008.

[4] R. M. Bittman, J. P. Romano, C. Vallarino and M. Wolf. Optimal testing of multiple hypotheses with common effect direction. *Biometrika*, **96**, 399–410, 2009.

[5] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[6] M. H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

[7] D. B. Duncan. A Bayesian approach to multiple comparisons. *Technometrics*, **7**, 171–222, 1965.

[8] R. Engelking. *General Topology*. Sigma series in pure mathematics. Heldermann, 1989.

[9] A. Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, **17**, 347–388, 2008.

[10] H. Finner and V. Gontscharuk. Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 1031–1048, 2009.

[11] H. Finner and K. Strassburger. The partitioning principle: a powerful tool in multiple decision theory. *Annals of Statistics*, **30**, 1194–1213, 2002.

[12] K. R. Gabriel. Simultaneous test procedures - some theory of multiple comparisons. *The Annals of Mathematical Statistics*, **41**, 224–250, 1969.

[13] Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, Inc., 1987.

[14] G Hommel. Multiple test procedures for arbitrary dependence structures. *Metrika*, **33**, 321–336, 1986.

[15] G. Hommel and F. Bretz. Aesthetics and power considerations in multiple testing – a contradiction? *Biometrical Journal*, **50**, 657–666, 2008.

[16] G. Hommel, F. Bretz and W. Maurer. Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine*, **26**, 4063–4073, 2007.
[17] R. Izbicki. Classes de testes de hipóteses (in Portuguese). Master's Thesis, University of São Paulo, 2010.
[18] R. Izbicki, V. Fossaluza, A. G. Hounie, E. Y. Nakano and C. A. de B. Pereira. Testing allele homogeneity: the problem of nested hypotheses. *BMC Genetics*, **13**, 1–11, 2012.
[19] E. T. Jaynes. Confidence intervals vs bayesian intervals. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, W. Harper and C. Hookers, eds, Springer Netherlands, pp. 175–257, 1976.
[20] H. Jeffreys. *Theory of Probability*. Cambridge University Press, 1939.
[21] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*, 6th edn. Prentice Hall, 2007.
[22] V. E. Johnson. Uniformly most powerful bayesian tests. *The Annals of Statistics*, **41**, 1716–1741, 2013.
[23] H. E. Kyburg, Jr. *Probability and the Logic of Rational Belief*. Wesleyan University Press, 1961.
[24] M. Lavine and M. Schervish. Bayes factors: what they are and what they are not. *The American Statistician*, **53**, 119–122, 1999.
[25] E. L. Lehmann. A theory of some multiple decision problems, i. *The Annals of Mathematical Statistics*, **28**, 1–25, 1957.
[26] E. L. Lehmann. A theory of some multiple decision problems, ii. *The Annals of Mathematical Statistics*, **28**, 547–572, 1957.
[27] I. Levi. *Gambling with Truth: An Essay on Induction and the Aims of Science*. MIT Press Classic, 1967.
[28] S. K. Lin, C. K. Chen, D. Ball, H. C. Liu and E. W. Loh. Gender-specific contribution of the $gaba_a$ subunit genes on 5q33 in methamphetamine use disorder. *Pharmacogenomics Journal*, **3**, 349–355, 2003.
[29] D. V. Lindley. *Introduction to Probability and Statistics from Bayesian Viewpoint, Part 2*. Cambridge University Press, 1965.
[30] M. R. Madruga, L. G. Esteves and S. Wechsler. On the Bayesianity of Pereira-Stern tests. *Test*, **10**, 291–299, 2001.
[31] R. Marcus, P. Eric and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660, 1976.
[32] D. G. Mayo and A. Spanos. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science*, **57**, 323–357, 2006.
[33] A. G. Patriota. A classical measure of evidence for general null hypotheses. *Fuzzy Sets and Systems*, **233**, 74–88, 2013.
[34] C. A. de B. Pereira and J. M. Stern. Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy*, **1**, 99–110, 1999.
[35] E. Raviv. On p-value. http://eranraviv.com/blog/on-p-value/. Accessed 26 March 2015.
[36] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
[37] C. Robert. *The Bayesian choice: From Decision-theoretic Foundations to Computational Implementation*. 2nd edn, Springer, 2007.
[38] J. P. Romano, A. M. Shaikh and M. Wolf. Consonance and the closure method in multiple testing. *The International Journal of Biostatistics*, **7**, 1–25, 2011.
[39] M. Rosenblum. Tests that reject at least one subpopulation null hypothesis after rejecting for overall population. *Johns Hopkins University, Dept. of Biostatistics Working Papers.*, pp. 347–88, 2012.

[40] D. Russell. Equal employment opportunity commission v. federal reserve bank of richmond. In *698 Federal Reporter 2d Series*, pp. 633–675. United States Court of Appeals, Fourth Circuit, 1983.
[41] E. Schechter. *Handbook of Analysis and Its Foundations*. Elsevier Science, 1996.
[42] M. J. Schervish. P values: what they are and what they are not. *The American Statistician*, **50**, 203–206, 1996.
[43] M. J Schervish. *Theory of Statistics*. Springer, 1997.
[44] J. P. Shaffer. Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, **81**, 826–831, 1986.
[45] J. P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, **46**, 561–584, 1995.
[46] G. M. Silva. Monotonicidade em testes de hipóteses (in Portuguese). Master's Thesis, University of São Paulo, 2010.
[47] E. Sonnemann. *Allgemeine Lösungen multipler Testprobleme*. Institut für mathematische Statistik und Versicherung, 1982.
[48] E. Sonnemann. General solutions to multiple testing problems. *Biometrical Journal*, **50**, 641–656, 2008.
[49] E. Sonnemann and H. Finner. Vollständigkeitssätze für multiple testprobleme. In *Multiple Hypothesenprüfung*, P. Bauer, G. Hommel and E. Sonnemann, editors, pp. 121–135. Springer, 1988.
[50] G. Stefansson, W Kim and J. C. Hsu. On confidence sets in multiple comparisons. *Statistical Decision Theory and Related Topics IV*, **2**, 89–104, 1988.
[51] J. M. Stern. Constructive verification, empirical induction, and falibilist deduction: a threefold contrast. *Information*, **2**, 635–650, 2011.
[52] A. R Templeton. Coherent and incoherent inference in phylogeography and human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 6376–81, 2010.
[53] R. A. Waller and D. B. Duncan. A Bayes rule for the symmetric multiple comparisons problem. *Journal of the American Statistical Association*, **64**, 1484–1503, 1969.
[54] H. Zhao, B. Wang and X. Cui. General solutions to consistency problems in multiple hypothesis testing. *Biometrical Journal*, **52**, 735–746, 2010.

Received 21 December 2014

Appendix: Proof of Part 2 of Theorem 3.11

Suppose that the relative losses are proper but that $\exists A, B \in \sigma(\Theta)$, with $A \subset B$, and $\exists \theta_1 \in A$ e $\theta_2 \in B^c$ with $L_A(0,\theta) - L_A(1,\theta) < L_B(0,\theta) - L_B(1,\theta)$, $\theta \in \{\theta_1, \theta_2\}$. As the relative losses are proper, we have

$$L_A(0,\theta_1) - L_A(1,\theta_1) < L_B(0,\theta_1) - L_B(1,\theta_1) \leq 0 \leq L_A(0,\theta_2) - L_A(1,\theta_2) < L_B(0,\theta_2) - L_B(1,\theta_2).$$

Consider that $x \in \mathcal{X}$ is observed and that the prior on θ is

$$\mathbb{P}(\{\theta_1\}) = \frac{pL_x(\theta_2)}{pL_x(\theta_2) + (1-p)L_x(\theta_1)} \text{ and } \mathbb{P}(\{\theta_2\}) = 1 - \mathbb{P}(\{\theta_1\}),$$

so that the posterior probability of θ given x, μ_x, is $\mu_x(\{\theta_1\}) = p$ e $\mu_x(\{\theta_2\}) = 1-p$, $0 < p < 1$. We divide the proof in four cases:

- If $L_B(0,\theta_1)-L_B(1,\theta_1)<0<L_A(0,\theta_2)-L_A(1,\theta_2)$, from the hypotheses it holds that

$$L_A(0,\theta_1)-L_A(1,\theta_1)<L_B(0,\theta_1)-L_B(1,\theta_1)<0<L_A(0,\theta_2)-L_A(1,\theta_2)<L_B(0,\theta_2)-L_B(1,\theta_2).$$

Hence,

$$x_1 \stackrel{\text{def}}{=} \frac{|L_A(0,\theta_2)-L_A(1,\theta_2)|}{|L_A(0,\theta_1)-L_A(1,\theta_1)|} < \frac{|L_B(0,\theta_2)-L_B(1,\theta_2)|}{|L_B(0,\theta_1)-L_B(1,\theta_1)|} \stackrel{\text{def}}{=} x_2.$$

If $p\in(0,1)$ is such that $x_1<\frac{p}{1-p}<x_2$,

$$\int_\Theta [L_A(0,\theta)-L_A(1,\theta)]d\mu_x(\theta)=[L_A(0,\theta_1)-L_A(1,\theta_1)]p+[L_A(0,\theta_2)-L_A(1,\theta_2)](1-p)<0<$$
$$<[L_B(0,\theta_1)-L_B(1,\theta_1)]p+[L_B(0,\theta_2)-L_B(1,\theta_2)](1-p)=\int_\Theta [L_B(0,\theta)-L_B(1,\theta)]d\mu_x(\theta).$$

Hence, $\mathcal{L}(A)(x)=0$, but $\mathcal{L}(B)(x)=1$.

- If $L_B(0,\theta_1)-L_B(1,\theta_1)=0<L_A(0,\theta_2)-L_A(1,\theta_2)$, we have

$$L_A(0,\theta_1)-L_A(1,\theta_1)<L_B(0,\theta_1)-L_B(1,\theta_1)=0<L_A(0,\theta_2)-L_A(1,\theta_2)<L_B(0,\theta_2)-L_B(1,\theta_2).$$

If $p\in(0,1)$ is such that

$$\frac{|L_A(0,\theta_2)-L_A(1,\theta_2)|}{|L_A(0,\theta_1)-L_A(1,\theta_1)|}<\frac{p}{1-p},$$

$$\int_\Theta [L_A(0,\theta)-L_A(1,\theta)]d\mu_x(\theta)=[L_A(0,\theta_1)-L_A(1,\theta_1)]p+[L_A(0,\theta_2)-L_A(1,\theta_2)](1-p)<0<$$
$$<[L_B(0,\theta_1)-L_B(1,\theta_1)]p+[L_B(0,\theta_2)-L_B(1,\theta_2)](1-p)=\int_\Theta [L_B(0,\theta)-L_B(1,\theta)]d\mu_x(\theta).$$

Hence, $\mathcal{L}(A)(x)=0$, but $\mathcal{L}(B)(x)=1$.

- If $L_B(0,\theta_1)-L_B(1,\theta_1)<0=L_A(0,\theta_2)-L_A(1,\theta_2)$, we have

$$L_A(0,\theta_1)-L_A(1,\theta_1)<L_B(0,\theta_1)-L_B(1,\theta_1)<0=L_A(0,\theta_2)-L_A(1,\theta_2)<L_B(0,\theta_2)-L_B(1,\theta_2).$$

If $p\in(0,1)$ is such that

$$\frac{p}{1-p}<\frac{|L_B(0,\theta_2)-L_B(1,\theta_2)|}{|L_B(0,\theta_1)-L_B(1,\theta_1)|},$$

we have

$$\int_\Theta [L_A(0,\theta)-L_A(1,\theta)]d\mu_x(\theta)=[L_A(0,\theta_1)-L_A(1,\theta_1)]p+[L_A(0,\theta_2)-L_A(1,\theta_2)](1-p)<0<$$
$$<[L_B(0,\theta_1)-L_B(1,\theta_1)]p+[L_B(0,\theta_2)-L_B(1,\theta_2)](1-p)=\int_\Theta [L_B(0,\theta)-L_B(1,\theta)]d\mu_x(\theta).$$

Hence, $\mathcal{L}(A)(x)=0$, but $\mathcal{L}(B)(x)=1$.

- If $L_B(0,\theta_1)-L_B(1,\theta_1)=0=L_A(0,\theta_2)-L_A(1,\theta_2)$, we have

$$L_A(0,\theta_1)-L_A(1,\theta_1)<L_B(0,\theta_1)-L_B(1,\theta_1)=0=L_A(0,\theta_2)-L_A(1,\theta_2)<L_B(0,\theta_2)-L_B(1,\theta_2).$$

For every $p\in(0,1)$,

$$\int_\Theta [L_A(0,\theta)-L_A(1,\theta)]d\mu_x(\theta)=[L_A(0,\theta_1)-L_A(1,\theta_1)]p+[L_A(0,\theta_2)-L_A(1,\theta_2)](1-p)<0<$$

$$<[L_B(0,\theta_1)-L_B(1,\theta_1)]p+[L_B(0,\theta_2)-L_B(1,\theta_2)](1-p)=\int_\Theta [L_B(0,\theta)-L_B(1,\theta)]d\mu_x(\theta).$$

Hence, $\mathcal{L}(A)(x)=0$, but $\mathcal{L}(B)(x)=1$.