

# Approaches for the joint evaluation of hypothesis tests: classical testing, Bayes testing, and joint confirmation

Robert M. Kunst  
University of Vienna

July 12, 2005

## **Abstract**

The occurrence of decision problems with changing roles of null and alternative hypotheses has increased interest in extending the classical hypothesis testing setup. Particularly, confirmation analysis has been in the focus of some recent contributions in econometrics. Within a general framework for decision problems, we demonstrate that classical testing and confirmation analysis are just two special cases.

Differences across the three approaches—traditional classical testing, Bayes testing, joint confirmation—are highlighted for a popular testing problem. A decision is searched for the existence of a unit root in a time-series process on the basis of two tests. One of them has the existence of a unit root as its null hypothesis and its non-existence as its alternative, while the roles of null and alternative are reversed for the second hypothesis test.

The aim of this contribution is, for one, to interpret the joint confirmation approach against the background of classical and Bayes testing and, secondly, to introduce decision-contours charts as convenient summaries of Bayes tests that build on two test statistics with different informational contents.

# 1 Introduction

The occurrence of decision problems with changing roles of null and alternative hypotheses has increased the interest in extensions of the classical hypothesis testing setup. Particularly, confirmation analysis has been in the focus of some recent contributions in econometrics (see DHRYMES, 1998, KEBLOWSKI AND WELFE, 2004, among others). This paper targets a general framework for decision problems, within which classical testing and confirmation analysis are just two special cases. We make the point that, in many situations, Bayes testing may be preferable to local principles.

The aim of this contribution is, for one, to interpret the joint confirmation approach against the background of classical and Bayes testing and, secondly, to introduce decision-contours charts as convenient summaries of Bayes tests that build on two test statistics with different informational contents.

In three sections, this paper builds up an appropriate framework for the general statistical decision problem. Section 2 reviews the problem as it is viewed in the more classical (see, for example, LEHMANN, 1959) or in the more Bayesian tradition (see, for example, FERGUSON, 1967, or PRATT *et al.*, 1995). Section 3 studies the Bayesian solution in more detail. Section 4 focuses on the more recent suggestion of ‘joint confirmation’, as it was used by CHAREMZA AND SYCZEWSKA (1998, CS), and show how it can be embedded into the general framework.

Section 5 summarizes the differences and similarities among the three approaches in a general framework. In Section 6, we demonstrate the differences across the approaches—traditional classical testing, Bayes testing, and joint confirmation—against the background of a statistical testing problem that was considered by CS and KEBLOWSKI AND WELFE (2004). A decision is searched for the existence of a unit root in a time-series process on the basis of two tests. One of them has the existence of a unit root as its null hypothesis and its non-existence as its alternative, while the roles of null and alternative are reversed for the second hypothesis test. The example is meant to highlight specific features and differences of the three approaches rather than to provide Bayesian decision points for empirical applications. Section 7 concludes.

## 2 The general decision problem

In any statistical decision problem, sets of probability laws constitute the possible choices. While it is of some interest to achieve generality beyond the typical case, it is advisable to start the presentation under the usual restrictive assumptions. In particular, we assume that the statistician searches for a decision between only two choice sets, which can be named the *null hypothesis* and the *alternative hypothesis*. It is also convenient that probability laws can be characterized by density functions and that these densities are continuous w.r.t. a common measure.

It is also not very restrictive to assume that both hypotheses can be expressed by a common parameter space  $\Theta$ , such that  $\Theta = \Theta_0 \cup \Theta_1$  and  $\Theta_0 \cap \Theta_1 = \emptyset$ . If the typical element  $\theta \in \Theta$  is a member of  $\Theta_0$ , ‘the null hypothesis is correct’, while for  $\theta \in \Theta_1$  the ‘alternative is correct’. In ‘parametric’ testing problems, the space  $\Theta$  has a finite dimension, which allows viewing  $\Theta_0$  and  $\Theta_1$  as being isomorphic to subsets of  $\mathbb{R}^k$  for some  $k$ . Truly non-parametric testing problems are rare. Rather, many testing problems are semiparametric in the sense that  $\Theta$  is indeed infinite dimensional but finite-dimensional projections of  $\theta$  split  $\Theta$  into a finite-dimensional set of classes. If the finite-dimensional projection of  $\theta$  is observed,  $\theta$  can be allotted to  $\Theta_0$  or  $\Theta_1$  with certainty.

**Example.** A variable  $X$  may be observed which is a realization of an unknown real-valued probability law with finite expectation.  $\Theta_0$  may consist of those probability laws that have  $\mathbb{E} X = 0$ , while  $\Theta_1$  may be defined by  $\mathbb{E} X \neq 0$ . Decision is searched for a one-dimensional parameter, while  $\Theta$  is infinite-dimensional.

A characteristic feature of statistical decision problems is that  $\theta$  is not observed. If  $\theta$  were observed, no problem arises at all and classification is perfect. This perfect case can be allotted a specific real number, for example zero. This value of zero corresponds to the usual concept of the *loss* incurred by a decision. Typically, a sample of observations for a random variable  $X$  is available to the statistician, where the probability law of the random variable is governed by the density  $f(\theta)$ . Instinctively, one may think that observing an infinite sequence of such observations will allow to determine  $\theta$  completely and therefore to attain the loss of zero that accrues from direct observation of  $\theta$ . Unfortunately, this is not always the case. Particularly for dependent observations, as they are common in time series, lack of ergodicity allows to construct simple examples where even a complete time series with infinitely many observations does not reveal the parameter of interest, which classifies  $\theta$  into  $\Theta_0$  or  $\Theta_1$ . It makes sense to rule out such oddities and to assume that a loss of zero *can* be attained in principle, using an infinite sequence of observations from  $X$ .

Typically, finite samples will not imply a loss of zero. Of course, if the null hypothesis is defined by probability laws on a specified bounded support, a single observation outside the bounds allows a zero-loss classification to  $\Theta_1$ . However, even in this simple constructed example no safe classification to  $\Theta_0$  is possible. In most real-life situations, there will be a certain probability of choosing  $\Theta_0$  or  $\Theta_1$  incorrectly. According to classical tradition, incorrect classification to  $\Theta_0$  is called a *type II error*, while incorrect classification to  $\Theta_1$  is called a *type I error*.

The above discussion implies that, at least usually, both the *type I* and the *type II* errors can be taken to zero probability, as the sample grows to infinity. Curiously, the most popular statistical decision tools, i.e. hypothesis tests with ‘fixed significance level’, do not serve this aim. These tests assume a setting where the probability of choosing  $\Theta_1$  while really  $\theta \in \Theta_0$  is bounded away from unity and can be made arbitrarily small by allowing for a larger probability of type II errors.

### 3 The Bayesian setup

The Bayesian setup to testing problems assumes weighting functions  $h_0$  and  $h_1$  on the respective parameter spaces  $\Theta_0$  and  $\Theta_1$ , which can be interpreted as probability densities. While a usual interpretation of  $h_0$  and  $h_1$  is that they represent *a priori* probabilities of parameter values, it is not necessary to adopt this interpretation for Bayes testing. If the sample space, for example  $\mathbb{R}^n$  for sample size  $n$ , is partitioned into two mutually exclusive subsets  $\Xi_0$  and  $\Xi_1$ , such that  $X \in \Xi_j$  implies deciding for  $\theta \in \Theta_j$ , the probability for a *type I* error is

$$\mathcal{P}_1(\theta) = \int_{\Xi_1} f_\theta(x) dx$$

for a given member  $\theta \in \Theta_0$ . The Bayes weighting scheme allows to evaluate

$$\mathcal{L}_0(h_0, \Xi_1) = \int_{\Theta_0} \int_{\Xi_1} f_\theta(x) dx h_0(\theta) d\theta$$

as a measure for the ‘average’ *type I* error involved in the decision. Conversely, the integral

$$\mathcal{L}_1(h_1, \Xi_0) = \int_{\Theta_1} \int_{\Xi_0} f_\theta(x) dx h_1(\theta) d\theta = \int_{\Theta_1} \mathcal{P}_0(\theta) h_1(\theta) d\theta$$

represents the ‘average’ *type II* error involved. A typical Bayesian view of the decision problem is to minimize

$$\begin{aligned} & g(\mathcal{L}_0(h_0, \Xi_1), \mathcal{L}_1(h_1, \Xi_0)) \\ &= g\left(\int_{\Theta_0} \int_{\Xi_1} f_\theta(x) dx h_0(\theta) d\theta, \int_{\Theta_1} \int_{\Xi_0} f_\theta(x) dx h_1(\theta) d\theta\right) \end{aligned}$$

in the space of possible partitions of the sample space, for a given function  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^+$ . The function  $g$  is designed to express the above-mentioned loss. Therefore,  $g(0, 0) = 0$  and monotonicity in both arguments are useful restrictions. If for any  $\theta \in \Theta_j$  no observed sample generated from that  $\theta$  implies the incorrect decision  $\Xi_k$  with  $k \neq j$ , both arguments are zero and the loss is zero. By construction,  $g(., .) = 0$  can also be attained if incorrect decisions occur for subsets  $\tilde{\Theta}_j \subset \Theta_j$  with  $h_j = 0$  or  $\int_{\tilde{\Theta}_j} h_j d\theta_j = 0$  only. Care must be taken not to define away problem areas of the parameter space by the choice of the weighting priors.

Classical testing usually proceeds under the assumption that  $\mathcal{P}_1(\theta)$  does not depend on  $\theta \in \Theta_0$  ‘too much’. The backdrop case is that  $\Theta_0$  is a singleton. In a slight extension of this case,  $\mathcal{P}_1(\theta)$  is assumed to be bounded for all  $\theta \in \Theta_0$ . Usually, a small bound is imposed, for example  $\alpha = 0.05$  or  $\alpha = 0.01$ . Usually, classical tests are constructed so that the bound of  $\alpha$  is actually attained for some

$\theta \in \Theta_0$ . Therefore, in classical testing the first integral will be less than  $\alpha$  for *any* weighting function  $h_0$ , although it will be ‘close’ to the bound  $\alpha$ . The textbook definition states that “ $\alpha = \sup \int_{\Xi_1} f_\theta(x) dx$  is the *size* or *significance level* of the test” [from SPANOS, p. 291]. In practice, tests often violate this condition. Apparently, then the size of the test is loosely defined as the exact value of  $\mathcal{P}_1(\theta)$  for some members  $\theta \in \Theta_0$ , which is not too much exceeded by other  $\theta \in \Theta_0$ .

By contrast, classical testing usually proceeds under the assumption that  $\mathcal{P}_0(\theta) = \int_{\Xi_0} f_\theta(x) dx$  depends critically on  $\theta \in \Theta_1$ . If  $\Theta_0$  is contained in the topological closure of  $\Theta_1$ , some  $\theta \in \Theta_1$  will imply integrals that come arbitrarily close to  $1 - \alpha$ , which is a relatively ‘large’ value. Other parameter values imply much smaller values of the *type II* error probability, and the weighted integral  $\mathcal{L}_1(h_1, \Xi_0)$  will typically be much smaller than  $1 - \alpha$  for most reasonable weighting functions  $h_1$ . However, it will always be possible to increase  $\mathcal{L}_1(h_1, \Xi_0)$ , such that it becomes arbitrarily close to  $1 - \alpha$ , simply by putting more weight on the values close to  $\Theta_0$ .

Classical statistics views a ‘good’ test as a decision procedure that has the prescribed value of  $\alpha$  as a maximum or upper bound for  $\mathcal{P}_1(\theta)$  with  $\theta \in \Theta_0$  and  $1 - \alpha$  as an upper bound for  $\mathcal{P}_0(\theta)$  with  $\theta \in \Theta_1$ . In applications, these conditions are not obeyed rigorously but they define what is known as an ‘unbiased’ test. Among unbiased tests, a test is viewed as being ‘better’ than a competing test if  $\mathcal{P}_0(\theta)$  is smaller for some values  $\theta \in \Theta_1$  and maybe equal for other values. It is not without interest that a potential gain relative to  $\alpha$  for some  $\theta \in \Theta_0$  is typically not seen as an advantageous feature of the procedure. Loosely, classical tests can be viewed as special Bayes tests with weighting functions that put all  $\Theta_0$  mass on the worst cases and most  $\Theta_1$  mass on values close to  $\Theta_0$ . This interpretation is confirmed by the definition of a ‘local maximum power’, which means that the probability of a *type II* error decreases fast close to the boundary in the parameter spaces.

For all but the most trivial decision problems, there is no test that minimizes  $\mathcal{P}_0(\theta)$  uniformly for all  $\theta \in \Theta_1$ . It is debatable whether minimization close to the boundary of  $\Theta_1$  versus  $\Theta_0$  should be the main concern. Assuming that  $\Theta$  can be metrified with a metric  $d$  and that  $d(\Theta_0, \theta)$  is unbounded for  $\theta \in \Theta_1$ , a useful requirement may be that  $\mathcal{P}_0(\theta)$  converge to 0 as  $d(\Theta_0, \theta) \rightarrow \infty$ . While this feature can indeed be attained in many statistical problems, it is usually not given much emphasis. In a sense, it is in the focus of ‘asymptotics’ where  $\mathcal{P}_0(\theta) \rightarrow 0$  defines the property of *consistency* as the sample space grows toward  $\mathbb{R}^N$ .

Once more, we note the remarkable fact that ‘consistency’ usually refers to  $\mathcal{P}_0(\theta)$  only. While  $\mathcal{P}_0(\theta) \rightarrow 0$  for  $\theta \in \Theta_1$  and  $\mathcal{P}_1(\theta) \rightarrow 0$  for  $\theta \in \Theta_0$  can be attained jointly for many actual decision problems, such that  $g(\mathcal{L}_0(h_0, \Xi_1), \mathcal{L}_1(h_1, \Xi_0)) \rightarrow 0$  for reasonable choices of  $h_0, h_1$ , few classical texts assign a name to this important property. Sometimes, it is referred to as *full consistency*. In a Bayesian interpretation, the classical approach is often viewed as allotting a

strong relative implicit weight to  $\Theta_0$  in smaller samples, which makes way to a strong weight on  $\Theta_1$  as the sample size grows. We note, however, that this ‘weight’ refers to the corresponding derivatives  $g_1$  and  $g_2$  of the  $g$  function, not to the weighting priors  $h_0$  and  $h_1$ .

## 4 Joint confirmation

Usually, the condition  $\sup_{\theta \in \Theta_0} \mathcal{P}_1(\theta) = \alpha$  is unable to determine uniquely a decomposition  $(\Xi_0, \Xi_1)$  of the sample space. Therefore, attention focuses on decompositions that are (1) based on test statistics  $\tau(X) : \mathbb{R}^n \rightarrow \mathbb{R}$  and are (2) minimizing  $\mathcal{P}_0(\theta)$  over  $\theta \in \Theta_1$  in some sense. For a large class of well-behaved statistical problems, the celebrated lemma of Neyman-Pearson guarantees that this problem has a simple solution, with  $\tau$  being a likelihood-ratio statistic and  $(\Xi_0, \Xi_1)$  being represented by intervals in the image space of  $\tau$ . Then, for example  $\Xi_0 = \tau^{-1}(-\infty, \tau_c]$  and  $\Xi_1 = \tau^{-1}(\tau_c, \infty)$ , and  $\tau_c$  is determined as some fractile of the distribution of  $\tau(X(\theta))$  for  $\theta \in \Theta_0$ , where we use  $X(\theta)$  for a sample that is generated from the distribution with the parameter  $\theta$ . While, in many more realistic problems, the property  $\theta \in \Theta_0$  does not uniquely define the distribution of  $\tau(X)$ , the framework remains operational for a large amount of interesting problems.

The facts that minimization of  $\mathcal{P}_0(\theta)$  is not possible uniformly over  $\theta \in \Theta_1$  and that the conditions of the Neyman-Pearson lemma are often fulfilled for subsets of the hypotheses  $\Theta_0$  and  $\Theta_1$  only, may suggest to consider more than one test statistic simultaneously. For example,  $\tau_1$  may minimize misclassification risk for a different subset of  $\Theta_1$  than its ‘rival statistic’  $\tau_2$ , while  $\mathcal{P}_0(\theta)$  may be nearly identical for an area ‘close’ to  $\Theta_0$ . Then, one may consider  $(\tau_1, \tau_2) : \mathbb{R}^n \rightarrow \mathbb{R}^2$  and base the decomposition  $(\Xi_0, \Xi_1)$  on bivariate intervals. It is plausible to choose

$$\begin{aligned} (\tau_1, \tau_2)^{-1}((-\infty, \tau_{c1}] \times (-\infty, \tau_{c2}]) &\subset \Xi_0, \\ (\tau_1, \tau_2)^{-1}([\tau_{c1}, \infty) \times ([\tau_{c2}, \infty)) &\subset \Xi_1. \end{aligned}$$

However, it is unclear how to allot the remaining parts of the sample space, where the two statistics seemingly point to different conclusions. Typically, allotting these parts to  $\Xi_0$  results in ‘low power’, while allotting them to  $\Xi_1$  violates the ‘risk level’ condition. An interesting empirical solution is to ‘estimate’ the parameter  $\theta$  and thus to approximately determine whether  $\tau_1$  or  $\tau_2$  is the ‘better’ test and to prefer the decision implied by one of the two statistics accordingly. Then, however, the test decision is essentially univariate again.

**Example.** In a linear regression model, a coefficient is tested for equality to a given value, under the danger of serial correlation of the errors. It used to be a common practice to estimate the model by least squares and to base the decision on the  $t$ -statistic, i.e. the Neyman-Pearson statistic for the model that assumes uncorrelated errors, unless a residual statistic indicates serious autocorrelation. In

that case, the model was re-estimated with a Cochrane-Orcutt correction and an accordingly adjusted  $t$ -statistic was used. While this practice is frowned upon by current econometrics, its idea is plausible: choosing the version of the  $t$ -statistic that is supposedly ‘better’ in certain areas of a parameter space that contains the regression coefficients and the error correlation structure, among possibly various other parameters.

At least to a classical statistician, the situation is further complicated if  $\Theta_0$  corresponds to the *null* hypothesis for a Neyman-Pearson construction of a test statistic  $\tau_1$ , while  $\Theta_1$  corresponds to the *null* hypothesis for another test statistic  $\tau_2$ . It is obvious that this is impossible in standard situations where  $\Theta = \Theta_0 \cup \Theta_1 \subset \mathbb{R}^k$  and  $\Theta_0$  is defined within  $\Theta$  by  $r < k$  equality restrictions. Then,  $\Theta_0$  will always be the null hypothesis. It may be, however, that  $\Theta_0$  and  $\Theta_1$  are ‘informally’ identified with two distinctive restrictions within  $\Theta$ , such that  $\Theta$  is a strict superset of  $\Theta_0 \cup \Theta_1$ . Again informally, the remaining portion  $\Theta \setminus (\Theta_0 \cup \Theta_1)$  is allotted to  $\Theta_1$  for the construction of  $\tau_1$ , while it is allotted to  $\Theta_0$  for the construction of  $\tau_2$ .

**Example.** It is of interest whether an observed effect is due to an observed factor A or to another observed factor B. The formulation of the real-life problem indicates that one, in principle, does not allow for the possibilities that the effect (1) does not exist or (2) is due to neither of the two causes or (3) that it is caused by both. Then,  $\tau_1$  is obtained from restricting the influence of factor B at zero, while  $\tau_2$  is obtained by restricting factor A. As long as  $\tau_1$  ‘rejects’ and  $\tau_2$  ‘accepts’ or *vice versa*, the decision is clear, while joint acceptance and joint rejection will cause trouble.

We note once more that the problem has a relatively simple solution in Bayesian statistics. The admissible parameter space is redefined as  $\Theta_0 \cup \Theta_1$  and the remainder is *a priori* excluded, according to the informal statement of the decision problem. After fixing weight functions  $h_0$  and  $h_1$  on the hypotheses and a loss criterion  $g$ , the decision problem can be subjected to computer power and yields a solution that is optimal within the pre-defined set of admissible decompositions  $(\Xi_0, \Xi_1)$ . For example, one may restrict attention to decompositions that are based on the test statistics  $(\tau_1, \tau_2)$  and on bivariate intervals.

Within classical statistics, a different suggestion can be found in the literature, which is however not used widely and apparently has not been fully accepted yet in the research community. It bears the name of ‘joint confirmation hypothesis’ or also ‘confirmatory analysis’. According to CS, one targets a risk PJC, which is defined as deciding for  $\Theta_1$ , given the validity of  $\Theta_1$ . This is indeed just  $\mathcal{P}_1(\theta)$  for  $\theta \in \Theta_1$ . Since  $\mathcal{P}_0(\theta) = 1 - \mathcal{P}_1(\theta)$ , the PJC simply corresponds to one of the error integrals that were defined above. The error integral  $\mathcal{P}_0(\theta)$  is evaluated for some  $\theta$ , which are members of the *alternative* for the test construction  $\tau_1$  and members of the *null hypothesis* for the construction of  $\tau_2$ . Therefore,  $\mathcal{P}_0(\theta)$  expresses the probability that  $\tau_1$  would wrongly accept its null and  $\tau_2$  would correctly accept its null if the tests were used individually. If  $(\tau_1, \tau_2)$  is used jointly, it is only the

probability of an incorrect decision for some given  $\theta \in \Theta_1$ .

Usually, there is a manifold of pairs  $(\tau_{a1}, \tau_{a2})$  such that  $(\tau_1, \tau_2) = (\tau_{a1}, \tau_{a2})$  implies the condition  $\mathcal{P}_0(\theta) = 1 - \alpha$  for a given  $\alpha$ . Among them, joint confirmation selects critical points  $(\tau_{c1}, \tau_{c2})$  by the condition that  $\mathcal{P}_0(\theta)$  coincide for the two component tests that build on individual  $\tau_j$  and corresponding  $\tau_{cj}$ . While this superficially looks like a Bayesian critical point, where the probabilities of  $\Theta_1$  and  $\Theta_2$  coincide, no probability of hypotheses is used, as the whole procedure is built in the classical way, where hypotheses do not have probabilities. It is conceded that there is an informal interpretation of a Bayesian mechanism, where  $p$ -values are interpreted as such probabilities. However, a Bayes test would determine critical points by comparing probabilities for  $\Theta_0$  and  $\Theta_1$ , not two measures of probability for  $\Theta_1$ . An apparent advantage of joint confirmation is, however, that it avoids the Bayesian construction of weighting functions  $h_0$  and  $h_1$ .

## 5 An instructive comparison of three methods

For an instructive comparison across the methods, first consider Figure 1. For the sake of an example, we consider a situation where  $\Theta_0$  corresponds to the null of the test using  $\tau_1$  and to the alternative of the test using  $\tau_2$ . Both tests have their rejection regions to the ‘left’. It is convenient to ‘code’ both tests in terms of their respective null distributions. In other words, the diagram is not drawn in the original coordinates  $(\tau_1, \tau_2) \in \mathbb{R}^2$  but rather in the fractiles  $(F_1(\tau_1), F_2(\tau_2)) \in [0, 1]^2$  for distribution functions  $F_1$  and  $F_2$ . Because distribution functions are monotonous transforms, the information remains identical for any such functions. However, the interpretation of the diagrams improves if  $F_j$  corresponds to the ‘null distributions’ of  $\tau_j$ , assuming that a distribution of  $\tau_j$  under its null hypothesis is (approximately) unique. In short, we label the axes simply by  $F(\tau_1)$  and  $F(\tau_2)$ . Then, if the test using  $\tau_1$  rejects this is equivalent to a value of  $F_1(\tau_1)$  less than 0.05. Further, if the test using  $\tau_1$  rejects and the test using  $\tau_2$  does not, there is clear evidence in favor of hypothesis  $\Theta_1$ . Conversely, if the first test does not reject and the second test does so, we have clear evidence in favor of  $\Theta_0$ . If both tests accept or reject, the evidence remains unclear. This fact is expressed by leaving the north-east and south-west regions white. This is the typical result of combining two purely classical tests.

Next, consider Figure 2. It represents the results of joint confirmation. Rather than using the null distributions of the two test statistics  $\tau_1$  and  $\tau_2$ , we use here the null distribution of  $\tau_2$  but the alternative distribution of  $\tau_1$  and code the two test statistics accordingly. Usually, *an* alternative distribution does not exist, therefore one uses a representative element from the  $\tau_1$  alternative. If  $\tau_1$  rejects *and*  $\tau_2$  accepts, this is the ‘confirmation area’ of hypothesis  $\Theta_1$ . Its probability under the representative distribution from  $\Theta_1$  has a given probability  $\alpha$ . Along the main diagonal, individual rejection probabilities coincide, thus the corner

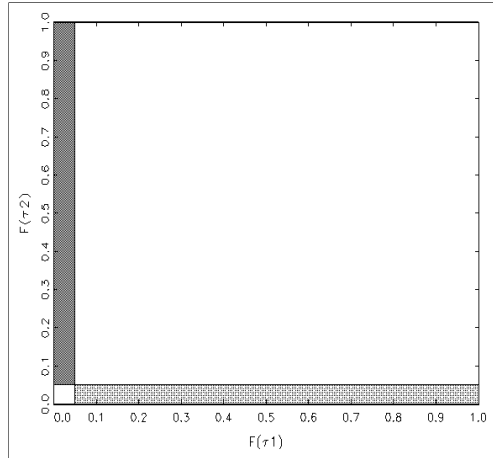


Figure 1: Classical decision following the joint application of two classical tests with switching null hypotheses. Axes are determined by the null distribution of  $\tau_1$  and the null distribution of  $\tau_2$ . Light gray area represents decisions in favor of  $\Theta_0$ , while the dark gray area corresponds to  $\Theta_1$ .

point is selected.

A possible interpretation of the method's focus on the north-west confirmation area is that the dark gray area favors  $\Theta_1$ , while the remaining area favors  $\Theta_0$ . The work of CS appears to support this interpretation by using a similar coloring of the four areas in a histogram plot. The interpretation is not coercive, however, and one may also follow the classical rule as in Figure 1. Then, joint confirmation becomes closer in spirit to reversing a classical test by replacing the original alternative by a point alternative, such that it becomes a convenient null. We refrain from this simplifying view, which is certainly invalid in the classical tradition, and view the joint confirmation decision according to 2. We note that the procedure is asymmetric in any case, as confirming  $\Theta_1$  leads to a different solution from confirming  $\Theta_0$ . The choice of confirmed hypothesis is not entirely clear. CS and KEBLOWSKI AND WELFE (2004) choose the null of the more popular component test.

A typical outcome of a Bayes test is depicted in Figure 3. As in the classical test in Figure 1, axes again correspond to respective null distributions. However, instead of a fixed null distribution  $F_j(x) = \int_{-\infty}^x f(z) dz$ , we now use a weighted average  $\int_{\Theta_j} f_{\theta}(z) h_j(\theta) d\theta$  of *all possible* null distributions. Then, a simulation with 50%  $\Theta_0$  and 50%  $\Theta_1$  distributions is conducted, where all kinds of representatives are drawn, according to weight functions  $h_0$  and  $h_1$ . Accordingly, a boundary can be drawn, where both hypotheses occur with the same frequency. Northwest of the boundary, which is sometimes called a *decision contour*, the

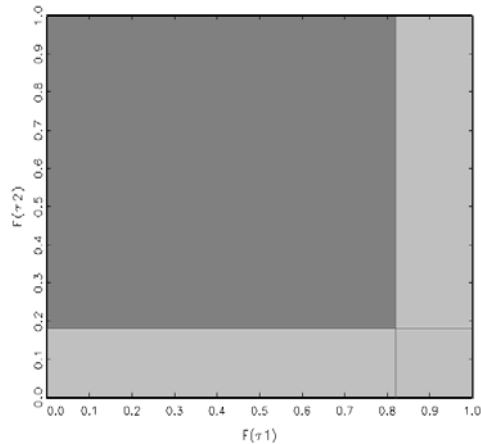


Figure 2: Joint-confirmation decision following the joint application of two classical tests with switching null hypotheses. Axes are determined by the null distribution of  $\tau_1$  and a representative alternative distribution of  $\tau_2$ . Light gray area represents decisions in favor of  $\Theta_0$ , while the dark gray area corresponds to  $\Theta_1$ .

hypothesis  $\Theta_0$  is preferred, while to the southeast the hypothesis  $\Theta_1$  is preferred. The decision rests on a much more informative basis than in the other approaches. However, the position of the curve is sensitive to the choice of  $h_0$  and  $h_1$ . In a fully Bayesian interpretation, the decision contour is defined as the set of all points  $\tau_c = (\tau_{c1}, \tau_{c2}) \in \mathbb{R}^2$  where  $P(\Theta_0|\tau_c) = P(\Theta_1|\tau_c)$ , if  $\Theta_j$  are given prior distributions of equal probability across hypotheses, i.e.  $P(\Theta_0) = P(\Theta_1)$  and the elements of the two hypotheses have prior probabilities according to the weight functions  $h_0$  and  $h_1$ . In the interpretation of the decision framework that we introduced in Section 2, the decision contour is simply the separating boundary of the two regions  $\Xi_0$  and  $\Xi_1$ , conditional on the restrictions that only such separations of the sample space are permitted that depend on the observed statistic  $\tau_c$  and on a loss function  $g(\cdot, \cdot)$  that gives equal weight to its two arguments, such as  $g(x, y) = x + y$ .

The choice of  $h_0$  and  $h_1$  is undoubtedly important for the Bayes test, as are all types of prior distributions for Bayesian estimation. There are several prescriptions for determining priors or ‘elicitations’ in the Bayesian literature. To some researchers, elicitation should reflect true prior beliefs, which however may differ subjectively and are maybe not good candidates for situations with strong uncertainty regarding the outcome. Other researchers suggest to standardize prior distributions and, consequently, weight functions according to some simple scheme. Particularly for Bayes testing aiming at deriving decision contours, it appears to be a good idea to keep the weight functions flat close to the rival hypothesis. Usually, the tail behavior of the weight functions has little impact

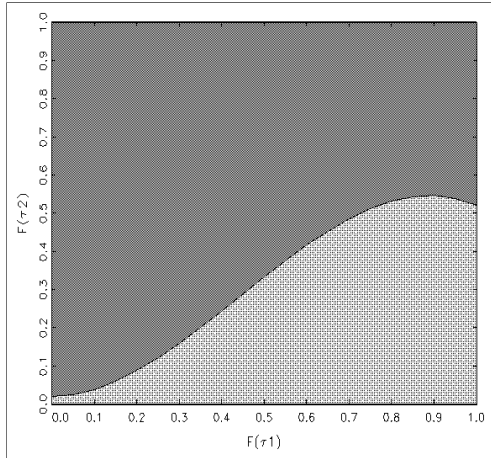


Figure 3: Bayes-test decision following the joint application of two classical tests with switching null hypotheses. Axes are determined by weighted averages of null distributions of  $\tau_1$  and  $\tau_2$ . Light gray area represents decisions in favor of  $\Theta_0$ , while the dark gray area corresponds to  $\Theta_1$ .

on the contours.

An important requirement is that the weighting priors are *exhaustive* in the sense that there is no  $\theta \in \Theta_j$  such that  $h_j(\theta) = 0$ . This ensures that any open environment within  $\Theta_j$  appears with positive weight in  $\mathcal{L}_0(h_0, \Xi_1)$  or  $\mathcal{L}_1(h_1, \Xi_0)$  and, consequently, that the loss  $g$  attains zero as  $n \rightarrow \infty$ . In technical terms, exhaustiveness means that any distribution within  $\Theta$  can be among the simulated draws.

Another important choice is the loss function  $g$ . The function  $g(x, y) = x + y$  corresponds to the Bayesian concept of allotting identical prior weights to the two hypotheses under consideration. In line with the scientific concept of unbiased opinion before conducting an experiment and in a search for ‘objectivity’, it appears difficult to accept loss functions such as  $g(x, y) = (1 - \kappa)x + \kappa y$  with  $\kappa \neq 1$ . These functions are sometimes used in the Bayesian literature (for example, see PRATT *et al.*, 1995) and may represent prior preferences for either of the two hypotheses. Interestingly, the classical tests with fixed significance levels can usually be interpreted as Bayes tests—with severe restrictions on the allowed decompositions  $(\Xi_0, \Xi_1)$ —with non-unit prior weights  $\kappa$ . Classical tests of this sort give preference to the null hypothesis in small samples and prefer the alternative for large samples. Again, seen from a Bayes-test viewpoint, it appears difficult to justify this traditional approach.

## 6 Testing for unit roots in time series

An important decision problem of time series analysis is to determine whether a given series stems from a stationary or a difference-stationary process. Stationary (or  $I(0)$ ) processes are characterized by the feature that the first two moments are constant in time, while difference-stationary (or  $I(1)$ ) processes are non-stationary but become stationary after first differencing. These two classes,  $I(0)$  and  $I(1)$ , are natural hypotheses for a decision problem. Various authors have provided different exact definitions of these properties, thereby usually restricting the space of considered processes. For example, instead of stationary processes one may focus attention on stationary ARMA processes, and instead of difference-stationary processes one may consider accumulated stationary ARMA processes.

This is approximately the framework of DICKEY AND FULLER (1979, DF) who introduced the still most popular testing procedure. Their null hypothesis  $\Theta_0$  contains finite-order autoregressive processes  $x_t = \sum_{j=1}^p \phi_j x_{t-j} + \varepsilon_t$ , formally written  $\phi(B)x_t = \varepsilon_t$  with white-noise errors  $\varepsilon_t$  and the property that  $\phi(1) = 0$ , while  $\phi(z) \neq 0$  for all  $|z| \leq 1$ , excepting the one unit root. We use the notation  $B$  for the lag operator  $BX_t = X_{t-1}$  and  $\phi(z) = 1 - \sum_{j=1}^p \phi_j z^j$  for general argument  $z$ . The corresponding alternative  $\Theta_1$  contains autoregressions with  $\phi(z) \neq 0$  for all  $|z| \leq 1$ . This is a semiparametric problem, as distributional properties of  $\varepsilon_t$  are not assumed, excepting the defining properties for the first two moments. In order to use asymptotic theorems, however, it was found convenient to impose some restrictions on some higher moments, typically of order three or four. We note that the interesting part of both hypotheses is fully parametric, and that both  $\Theta_0$  and  $\Theta_1$  can be viewed as equivalent to subspaces of  $\mathbb{R}^{\mathbb{N}}$ . In particular, this ‘interesting part’ of  $\Theta_0 \cup \Theta_1$  can be viewed as containing sequences of coefficients  $(\phi_j)_{j=0}^{\infty}$  with the property that  $\phi_0 = 1$  and  $\phi_j = 0$  for  $j > J$ , for some  $J$ . Choosing  $\Theta_0$  as the null hypothesis is the natural choice, as it is defined from the restriction  $1 - \sum_{j=1}^{\infty} \phi_j = 1 - \sum_{j=1}^J \phi_j = 0$  on the general space. Stated otherwise,  $\Theta_0$  has a ‘lower dimensionality’ than  $\Theta_1$ , even though both spaces have infinite dimension by construction.

In the form that is currently widely used, the test statistic is calculated as the  $t$ -statistic of  $a$  in the auxiliary regression

$$\Delta y_t = a y_{t-1} + \sum_{j=1}^{p-1} \xi_j \Delta y_{t-j} + u_t, \quad (1)$$

where  $\Delta$  denotes the first-difference operator  $1 - B$ ,  $(a, \xi_1, \dots, \xi_{p-1})'$  is a one-one transform of the coefficient sequence  $(\phi_1, \dots, \phi_p)'$ ,  $a = 0$  iff  $\phi(1) = 0$ ,  $p$  is either determined as a function of the sample size or by some empirical criterion aiming at  $u_t = \varepsilon_t$ , and  $u_t$  is the regression error. While the decision model was introduced in this purely autoregressive framework by DF—such that eventually

$p \geq J$  and  $u_t = \varepsilon_t$ —, it was extended to ARMA models by later authors. In other words, the test statistic continues to be useful if  $\varepsilon_t$  is MA rather than white noise, assuming some further restriction, such as excluding unit roots in the MA polynomial.

It is to be noted that the test suggested by DF is a purely classical test. The distribution of the test statistic in  $\Theta_0$  was tabulated for finite samples under special assumptions on the  $\varepsilon_t$  distribution, while the asymptotic distribution was later expressed by integrals over Gaussian continuous random processes (see, for example, DHRYMES, 1988). In finite samples,  $\mathcal{P}_1(\theta) = \alpha$  does not hold exactly for all  $\theta \in \Theta_0$ , and  $\mathcal{P}_0(\theta) \leq 1 - \alpha$  will not hold for all  $\theta \in \Theta_1$ . Straightforward application of the test procedure with fixed  $\alpha$  will result in  $\mathcal{P}_1(\theta) \rightarrow \alpha$  for  $\theta \in \Theta_0$  and  $\mathcal{P}_0(\theta) \rightarrow 0$  for  $\theta \in \Theta_1$  if  $n \rightarrow \infty$ . The Bayes loss  $g(\cdot, \cdot)$  will not decrease to 0 as  $n \rightarrow \infty$ .

Interestingly, several authors have studied the properties of the DF test *outside*  $\Theta_0 \cup \Theta_1$ . For example, it was found of interest to investigate the cases that the polynomial zero under I(1) has a multiplicity greater than one (see PANTULA, 1989) and that the processes have some simple features of non-stationarity under both I(0) and I(1) (see PERRON, 1989, and MADDALA AND KIM, 1998).

A different test for the unit roots problem is the KPSS test (after KWIATKOWSKI *et al.*, 1992). According to TANAKA (1996), its test statistic has the appealingly simple form

$$K = n^{-1} \frac{y' MCC' My}{y' My},$$

where  $y$  is the vector of data,  $M$  corrects for means or trends, and  $C$  is used to accumulate a series in the sense of the operator  $\Delta^{-1}$ . This version is correct for testing a null hypothesis that  $y$  is white noise against some alternative where  $y$  is a random walk, which is a most unlikely situation. If the null hypothesis is to contain more general stationary processes, KPSS suggest a non-parametric correction of the above statistic. The correction factor  $r$  is defined as  $r = \tilde{\sigma}_S^2 / \tilde{\sigma}_L^2$ , where  $\tilde{\sigma}_S^2$  is an estimate of the variance of a mean-corrected version of  $\Delta y$ ,  $\eta = \Delta y - m_{\Delta y}$  for  $m_{\Delta y} = (n-1)^{-1} \sum_{t=2}^n \Delta y_t$ , and  $\tilde{\sigma}_L^2$  is an estimate of the re-scaled zero-frequency spectrum of the same process. We follow the specification for these estimates as

$$\begin{aligned} \tilde{\sigma}_S^2 &= n^{-1} \sum_{t=2}^n \eta_t^2, \\ \tilde{\sigma}_L^2 &= \tilde{\sigma}_S^2 + 2n^{-1} \sum_{j=1}^l \left(1 - \frac{j}{l+1}\right) \sum_{t=j+2}^n \eta_t \eta_{t-j}. \end{aligned}$$

This is a Bartlett-type estimate of the frequency-zero spectrum. We follow one of the suggestions in the literature for specifying the upper bound  $l$  as the integer part of  $5\sqrt{n}/7$ . For a comparison, we consider the uncorrected version  $K$  as well as the corrected version  $\tilde{K} = rK$ . We start by presenting our results for  $K$ .

KEBLOWSKI AND WELFE (2004) give a 5% point, according to their construction, at -3.10 for the DF test and 0.42 for the KPSS test. Both tests have regions pointing to stationarity in the left tails of their null distribution. However, the classical interpretation differs across the two tests. For the DF test, integratedness is the null hypothesis, and rejection in the left tails indicates stationarity. For the KPSS test, stationarity is the null hypothesis, and rejection in the right tails indicates integratedness. Therefore, in classical testing, secure decisions are only taken in the areas  $[0, \alpha] \times [0, 1 - \alpha]$  and  $[\alpha, 1] \times [1 - \alpha, 1]$  of the coded null fractiles table that we used in Figure 1. By using the values -3.10 and 0.42, KEBLOWSKI AND WELFE determine an acceptance region for the unit root as  $(-3.10, \infty) \times (0.42, \infty)$ , thus suggesting to decide for the stationarity hypothesis in the remainder of  $\mathbb{R}^2$ . This situation could be drawn in a re-coded version just as in Figure 2, with the dark area now to the north-east (see also below, Figure 6)

In order to provide the Bayesian test solution, we consider parameterizing fully the two hypotheses  $\Theta_0$  and  $\Theta_1$ . The stationary processes are first-order autoregressions  $X_t = \phi X_{t-1} + \varepsilon_t$ , with  $\phi$  distributed uniformly on  $(-1, 1)$ . The integrated processes are simply random walks  $X_t = X_{t-1} + \varepsilon_t$ . The errors  $\varepsilon_t$  are drawn independently from a standard normal distribution. All trajectories are started from zero and have length  $n = 100$ . This design gives some unfair advantage to the DF test, as it corresponds to the design of DF. For the basic experiment, we stick to the simple design for simplicity. Note, however, that the thus defined weight functions  $h_0$  and  $h_1$  are exhaustive on a quite restricted set  $\Theta_0 \cup \Theta_1$  only.

After setting up the Bayesian weighting framework, defining implicitly  $h_0$  and  $h_1$ , trajectories are drawn and statistics, DF and KPSS, are being calculated for each trajectory. We chose the replication size of  $2 \times 10^6$ , meaning that  $10^6$  random-walk and  $10^6$  stationary trajectories were used. The DF statistics from the random walks and the KPSS statistics from the stationary autoregressions define the null distributions for further steps. While the DF distribution corresponds well to its tabulated and also to its asymptotic form, the KPSS null distribution cannot correspond to its theoretical one, which rests on white-noise trajectories. It is informative to draw the empirical distribution functions for null and alternative models. In the case of the DF statistic, the distribution functions have similar shapes and show a satisfactory discrepancy among them. This means that the average alternative model for the DF test may really correspond to white noise, as it should according to construction, while the drawn trajectories of course contain negatively correlated specimens as well as near random walks. We obtain a different picture for the KPSS statistic, where the alternative distribution has an almost uniform appearance. However, also for the KPSS statistic, there is a comforting difference in shape between the average null and the average alternative distribution.

The next step is setting up a grid in the fractile space. We chose a  $100 \times 100$

grid, which corresponds to 200 entries per bin, if the statistics had been jointly uniformly distributed. It turned out that many bins are indeed empty and that the smoothness of the boundary curve is not quite satisfactory. In this situation, increasing replications or kernel smoothing are possible options. One may also consider reducing the resolution of the grid.

Then, the grid is filled with the simulated statistics, the origin of which is known, according to whether they belong to  $\Theta_0$  or to  $\Theta_1$ . Finally, bins with a preponderance of stationary processes are interpreted as suggesting a decision in favor of  $\Theta_0$ , while other bins are allotted to  $\Theta_1$ . The separation of the sample space, as coded by null distributions of the two test statistics, into  $\Xi_0$  and  $\Xi_1$ , as it were, is shown in Figure 4. The boundary or *decision contour* runs south-east from the north-west corner. The large light gray area in the south-east marks test outcomes that were not observed in the simulation. The probability—in the Bayesian sense—of large or even average DF statistics *together* with small or even average KPSS statistics is very low. Supports remain unbounded, and an unreasonably large number of replications will succeed in determining the decision contour in the south-east fields. It is questionable whether the exercise is worth the computer time, as these values were not generated with a reason. They are simply very unusual in empirical practice. In order to corroborate this statement, one might require extending  $\Theta_0$  and  $\Theta_1$  to cover more general stationary and integrated processes. We will point at the consequences in some more sophisticated experiments below.

Note that the boundary is much more informative than any of the alternative decision concepts in classical statistics. For example, one sees that the DF statistics show a much sharper concentration under their stationary alternative than the KPSS statistics under their random-walk alternative. This feature may still reflect the missing correction for serial correlation in our KPSS version. One also obtains particular information on the behavior in the north-west corner of conflict, where DF statistics point to stationary behavior, while KPSS statistics indicate unit roots. The decision contour deviates from a straight  $(0, 1) - (1, 0)$  diagonal and appears to emphasize the role of the KPSS statistic in conflict situations. In summary, the graph allows a truly bivariate evaluation of the information provided by the two test statistic, which the classical approaches do not.

To assess the sensitivity of the boundary curve, we repeat the experiment with  $\Theta_0$  now consisting of moving-average processes with uniform  $\theta \in (-1, 1)$  and  $x_t = \varepsilon_t + \theta\varepsilon_{t-1}$ . Then,  $\Theta_1$  is not contained in the topological closure of  $\Theta_0$ , and discriminating the two hypotheses becomes easier. In Figure 5, we see that the boundary moves west, due to an extreme concentration of the distribution of simulated DF test statistics along the western border. The graph suggests to use the DF test at a significance level of 1%. In case of rejection, the observed time series should be stationary. In case of no rejection, the KPSS statistic will be ‘large’ and the time series will stem from an integrated process. We note that

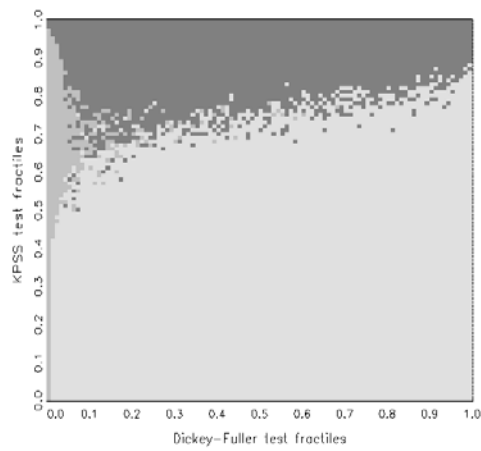


Figure 4: Bayes-test decision following the joint application of the Dickey-Fuller and KPSS tests with switching null hypotheses. Sample size is  $n = 100$ . Axes are determined by weighted averages of null distributions of the Dickey-Fuller and KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

these recommendations only make sense if we know that any potential stationary process is first-order moving-average, which is a most unlikely situation.

In order for the method to be useful to empirical researchers, some standardization of flexible features like  $h_0$  and  $h_1$  will be necessary. While such a standardization of weighting priors will hardly satisfy the convinced Bayesian, it appears to be a possibility and should correspond to the ubiquitous 5% of classical statistics. For example, in this example the design of  $h_0$  in Figure 5 is unsatisfactory, while the one in Figure 4 is ‘better’.

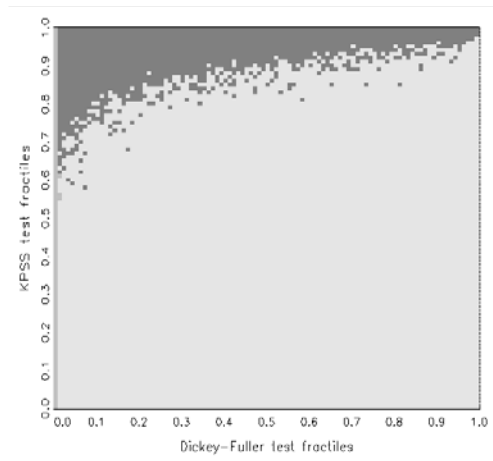


Figure 5: Bayes-test decision following the joint application of the Dickey-Fuller and KPSS tests with switching null hypotheses. Sample size is  $n = 100$ . Axes are determined by weighted averages of null distributions of the Dickey-Fuller and KPSS statistics. Stationary processes were generated from first-order moving-average processes with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

Now, we re-consider the critical point  $(-3.10, 0.42)$  that was provided by KE-BLOWSKI AND WELFE (2004) as solutions to the joint confirmation approach. In the coordinates of our diagram 4, this point is situated in the north-west corner approximately  $(0.03, 0.76)$ , not too distant from the three-countries corner. We note that the region  $(-3.10, \infty) \times (0.42, \infty)$  in coordinates of  $(\tau_1, \tau_2)$  and  $(0.03, 1) \times (0.76, 1)$  in the diagram coordinates belongs to the ‘confirmation area’ for  $I(1)$ . Thus, a large part of the confirmation area coincides with the hypothesis- $I(1)$  area of the Bayes test. Test outcomes in the north-west to the joint-confirmation point, however, would be classified differently. Test outcomes in the south-east are rare, and their classification is of little empirical relevance. The approximate coincidence of decisions is not a systematic property of the two

procedures, and it depends on the sample size. The choice of weighting functions for the Bayes procedures is not unique, hence the coordinate system looks differently for different  $h_0$  or  $h_1$ . For example, if draws for the  $I(0)$  hypothesis are restricted to the textbook case of white noise, instead of extending them to autoregressive or moving-average processes, the point  $(-3.10, 0.42)$  will appear as  $(0.03, 0.94)$ , which comes closer to the construction idea of joint confirmation. The decision map for the decision suggested by joint-confirmation analysis is given as Figure 6 in the coordinates of Figure 4.

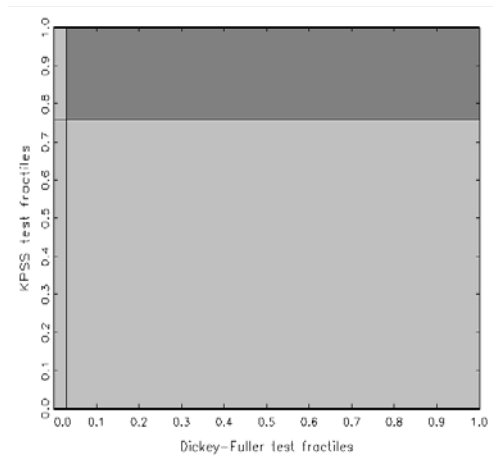


Figure 6: Joint-confirmation decision following the joint application of the Dickey-Fuller and KPSS tests with switching null hypotheses. The critical vertex was taken from the literature for the sample size  $n = 100$ . Axes are determined by weighted averages of null distributions of the Dickey-Fuller and KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration.

As the sample size increases, the distribution of test statistics in the Bayesian procedure converges to the western—for  $I(0)$ —and northern—for  $I(1)$ —borders, and the decision contour disappears in the north-west corner. By contrast, the decisions of classical and of joint-confirmation statistics rely on fixed respective critical points close to  $(0.05, 0.95)$  or  $(0.03, 0.76)$ . Convergence of joint-confirmation critical points to their limits is conveniently fast, as can be seen from the tables provided by CS or KEBLOWSKI AND WELFE (2004). Thus, for any simple weighting function  $g(., .)$ , such as  $g(x, y) = x + y$ , the loss incurred by the decision problem does not converge to zero for the classical methods, while it does so by construction for the Bayesian test design.

Figure 7 allows an impression of the influence of the sample size on decision contours. It shows an identical experiment to Figure 4, with the only difference

that  $n = 20$  instead of  $n = 100$ . In such small samples, the central area is still reasonably populated, and the decision contour spreads along the  $y = 1 - x$  diagonal, although with a slightly shifted position. Apparently, ‘rejection’ according to the KPSS statistic is given priority, such that the north-west corner is in the hands of the I(1) hypothesis.

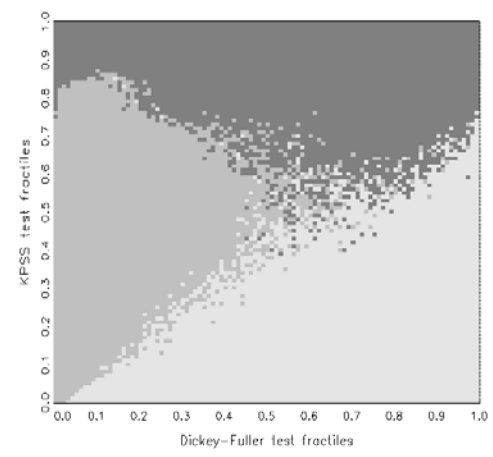


Figure 7: Bayes-test decision following the joint application of the Dickey-Fuller and KPSS tests with switching null hypotheses. Sample size is  $n = 20$ . Axes are determined by weighted averages of null distributions of the Dickey-Fuller and KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

If the uncorrected KPSS statistic  $K$  is replaced by the statistic  $\tilde{K}$ , which according to theoretical results is more appropriate in our design, we obtain the contour plot in Figure 8. It is obvious that the influence of the correction term is strong. Convergence of the finite-sample distribution of  $\tilde{K}$  to its limit is much slower than for the uncorrected  $K$ . The decision contour is now recognizable for a larger part of the diagram. The picture suggests to rely mainly on the decision suggested by the DF test. Time series with DF statistics that do not imply individual rejection are likely to be integrated, even when  $\tilde{K}$  is moderately low. The diagram does not imply that  $\tilde{K}$  is not a useful statistic, as the simulation design favors the parametric DF test. Neither does it imply that  $K$  should be used instead of  $\tilde{K}$ . However, it suggests that, assuming time series have actually been generated from first-order autoregressions, low  $\tilde{K}$  values observed together with inconspicuous DF values point to non-stationary generating processes. This information may be valuable and it is in outright contradiction to the classical *and* to the joint-confirmation approaches.

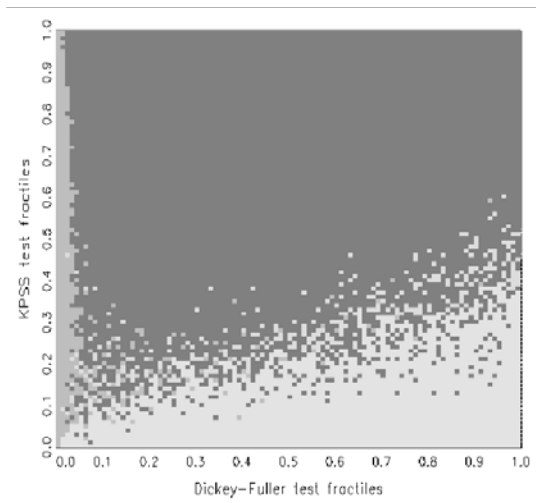


Figure 8: Bayes-test decision following the joint application of the Dickey-Fuller and corrected KPSS tests with switching null hypotheses. Sample size is  $n = 100$ . Axes are determined by weighted averages of null distributions of the Dickey-Fuller and corrected KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

It was pointed out before that the sampling design gives an advantage to the parametric procedure, as the generating model corresponds to the testing model of the DF test but not to the testing model of the KPSS test. In order to remove that advantage, we now consider a mixed generating design, where 50% of the  $I(0)$  and  $I(1)$  processes are generated as before. The remaining 50% are generated from sums of stationary first-order autoregressions and random walks in the form

$$\begin{aligned}y_t &= x_t + u_t, \\x_t &= x_{t-1} + \xi_t, \\u_t &= \phi u_{t-1} + \varepsilon_t.\end{aligned}$$

The stationary  $x_t$  process is generated as before, with  $\phi$  uniformly drawn from  $U(-1, 1)$  and  $\varepsilon_t$  drawn from  $N(0, 1)$ . The main difference is the  $N(0, \sigma^2)$  process  $\xi_t$ . For the  $I(0)$  hypothesis,  $\sigma^2 = 0$ , while for the  $I(1)$  hypothesis  $\sigma^2$  is drawn from a standard half-normal distribution. We note that the  $I(0)$  design is the same as before, such that 50% of all processes are generated from that model, while 25% obey the pure random walk null of the DF model and 25% obey the mixed state-space model that underlies the KPSS test. A similar concept was adopted by KUNST AND REUTTER (2002) in their evaluation of statistics for decisions on seasonal behavior.

For this more complex simulation design, we obtain the contour plot in Figure 9. Apparently, the decision for the DF test now changes to a value that is much closer to the null median than to the lower-tail fractiles that were observed in other charts. However, we note that the null distribution of the DF test has now also been changed, due to a different generating model for  $I(1)$ , and the shift of the DF contour to the right is much less pronounced in real values than may be suggested by the chart. However, the KPSS test now shows its strength, particularly in the north-west corner. This implies that large values of the KPSS statistic now imply an  $I(1)$  decision, even when the DF test tells otherwise, which is good news for the supporters of the KPSS test.

We also tried to replace the Dickey-Fuller test by an ‘augmented’ variant in the same simulation design. The augmented Dickey-Fuller test uses a regression

$$\Delta y_t = \alpha y_{t-1} + \sum_{j=1}^p \gamma_j \Delta y_{t-j} + u_t,$$

thus reducing serial correlation in the errors  $u_t$  and a null distribution that comes closer to the one under the pure random-walk hypothesis. The literature recommends to determine  $p$  from the data, while we simply set  $p = 2$  for the experiment. The result is shown in Figure 10. Note that in this and the previous figure, axis ticks have been replaced by the empirical null fractiles. In Figure 10, these values are again close to the random-walk null and the decision contour moves back to the familiar shape. This indicates that Figure 9 is probably not representative

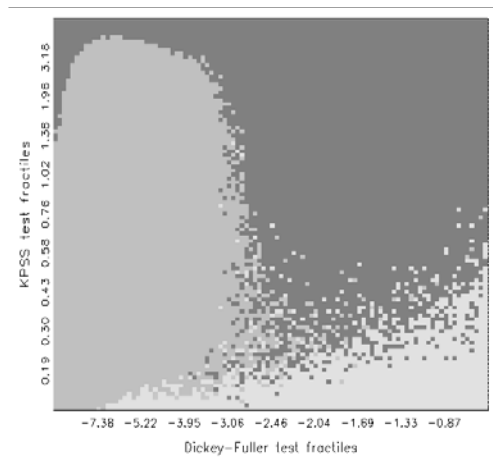


Figure 9: Bayes-test decision following the joint application of the Dickey-Fuller and corrected KPSS tests with switching null hypotheses. Sample size is  $n = 100$ . Sampling design corresponds to a mix of autoregressions and state-space processes. Axes are determined by weighted averages of null distributions of the Dickey-Fuller and corrected KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

for an advantage of KPSS testing but rather reflects an incorrect application of the DF test. We also experimented with variants of data-determined lag orders  $p$ , without any further important change in the overall shape.

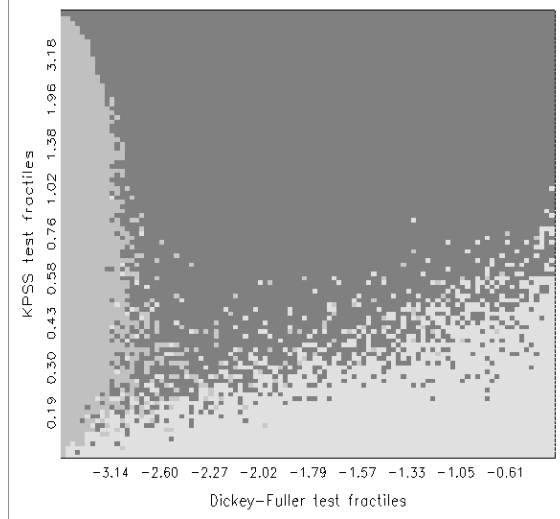


Figure 10: Bayes-test decision following the joint application of the augmented Dickey-Fuller and corrected KPSS tests with switching null hypotheses. Sample size is  $n = 100$ . Sampling design corresponds to a mix of autoregressions and state-space processes. Axes are determined by weighted averages of null distributions of the augmented Dickey-Fuller and corrected KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

## 7 Summary and conclusion

In this paper, we compare three approaches for obtaining a decision based on two univariate test statistics. These problems are common in statistical research, either because one wishes to consider two test statistics with similar properties but locally different power, or because one wishes to consider a very general maintained hypothesis, which is insufficiently represented by the null and alternative hypotheses that were used for constructing the test statistics. The latter case is of special interest, particularly if it involves an exchange of the roles of null and alternative hypotheses for the two test statistics.

The first approach, classical testing at a fixed significance level, is the most

common one but it is the least satisfactory. The approach entails inconsistent decisions as  $n \rightarrow \infty$  and provides no advice in cases of conflicting outcomes. A variant of this approach, classical testing at sample-dependent significance level, leads to consistent decisions but still fails with respect to the latter concern.

The second approach, joint confirmation analysis, provides an interesting solution to conflicting outcomes of component tests by adjusting the significance level to a point of one of the component alternatives by re-interpreting it as a null. Thereby, the approach succeeds in prescribing decisions for any outcome. Remaining within the boundaries of the traditional viewpoint, we feel that it does not account fully for the information provided by the pair of observed statistics. Another disadvantage may be its inherent asymmetry. While the asymmetric treatment of null and alternative in the classical paradigm may reflect the need to put subject-matter theories to a test, asymmetry is difficult to support if two tests are conducted with switching null and alternative hypotheses.

The third approach, Bayes testing with maps coded in the fractile space, succeeds in reaching fully consistent decisions and in automatically processing the provided information. Its drawback is its sensitivity to weight functions and its time-consuming simulation and evaluation. A major step in its widespread applicability could be the general standardization of weighting priors. Such standardization could provide a counterpart to the traditional classical significance levels.

The Bayes-test approach is also the most flexible one if it comes to extensions of the maintained hypotheses. Instead of crudely viewing cases outside of the maintained hypotheses in the construction stage of the utilized test statistics as belonging ‘rather’ to the null or alternative, one may simply re-do the simulations on extended parameter spaces or add a third hypothesis to the decision set. The approach remains valid for any finite set of decision alternatives, much beyond the traditional setup of ‘null’ and ‘alternative’. An application to a set of three alternatives was attempted by KUNST (2003). It is also conceivable to extend the approach in order to cover more than two univariate statistics, although the graphical interpretation would then be lost.

The particular application of the principles—discriminating stationarity from difference stationarity—was selected as it was intensely treated in time-series econometrics and constitutes one of the few examples for an application of the joint-confirmation approach in the literature. It should be noted that both the Bayes-test simulation design and the component tests can be improved. Recently, LEYBOURNE *et al.* (2005) found that a relatively simple modification of the Dickey-Fuller test statistic due to LEYBOURNE (1995) yields the most impressive power gains over several competing suggestions. The unusual null distribution of the Leybourne statistic is not a problem in the Bayes-test or in the joint-confirmation paradigm, as critical points or curves are determined by simulation anyway. Such refinements are a topic for further research with this particular focus. An evaluation of the loss  $g(\mathcal{L}_0(h_0, \Xi_1), \mathcal{L}_1(h_1, \Xi_0))$  can demon-

strate whether the gains in local power translate into global benefits for the underlying statistical decision problem.

With regard to a wider view on statistical decision problems, it appears that most current research focuses on improving the construction of test statistics, with the aim of improving on  $\mathcal{P}_0(\theta)$ . While the choice of test statistic is undoubtedly an important ingredient of the decision procedure and the whole chain from data to decision is certainly as strong as its weakest element, our research work emphasizes that the other elements, such as loss and weighting priors, may also deserve attention.

## References

- [1] CHAREMZA, W.W., AND E.M. SYCZEWSKA (1998) ‘Joint application of the Dickey-Fuller and KPSS tests’. *Economics Letters* **61**, 17–21.
- [2] DICKEY, D.A., AND W.A. FULLER (1979) ‘Distribution of the estimators for autoregressive time series with a unit root’. *Journal of the American Statistical Association* **74**, 427–431.
- [3] DHRYMES, P. (1998) *Time Series, Unit Roots, and Cointegration*. Academic Press.
- [4] FERGUSON, T.S. (1967) *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press.
- [5] KEBLOWSKI, P., AND A. WELFE (2004) ‘The ADF-KPSS test of the joint confirmation hypothesis of unit autoregressive root’. *Economics Letters* **85**, 257–263.
- [6] KUNST, R.M. (2003) ‘Decision maps for bivariate time series with potential threshold cointegration’. Working paper, Institute for Advanced Studies, Vienna.
- [7] KUNST, R.M., AND M. REUTTER (2002) ‘Decisions on seasonal unit roots’. *Journal of Statistical Computation and Simulation*. **72**, 403–418.
- [8] KWIATKOWSKI, D., PHILLIPS, P.C.B., SCHMIDT, P., AND Y. SHIN (1992) ‘Testing the null hypothesis of stationarity against the alternative of a unit root,’ *Journal of Econometrics* **54**, 159–178.
- [9] LEHMANN, E.L. (1959) *Testing Statistical Hypotheses*. Wiley.
- [10] LEYBOURNE, S. (1995) ‘Testing for unit roots using forward and reverse Dickey-Fuller regressions’. *Oxford Bulletin of Economics and Statistics* **57**, 559–571.

- [11] LEYBOURNE, S., KIM, T.H., AND P. NEWBOLD (2005) ‘Examination of some more powerful modifications of the Dickey-Fuller test’. *Journal of Time Series Analysis* **26**, 355–370.
- [12] MADDALA, G. S., AND I.M. KIM (1998) *Unit Roots, Cointegration, and Structural Change*. Cambridge University Press.
- [13] PANTULA, S.G. (1989) ‘Testing for unit roots in time series data’. *Econometric Theory* **5**, 256–271.
- [14] PERRON, P. (1989) ‘The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis’. *Econometrica* **57**, 1361-1401.
- [15] PRATT, J.W., RAIFFA, H., AND R. SCHLAIFER (1995) *Introduction to Statistical Decision Theory*. MIT Press.
- [16] SPANOS, A. (1986) *Statistical Foundations of Econometric Modelling*. Cambridge University Press.
- [17] TANAKA, K. (1996) *Time Series Analysis*. Wiley.