ÖSTERR. ZEITSCHRIFT FÜR STATISTIK
25. Jahrgang (1996), Heft 2, S. 7–32

Estimating Discrete Parameters: An Application to Cointegration and Unit Roots

Robert M. Kunst
Institute for Advanced Studies, Vienna
and
Johannes Kepler University Linz

Abstract: The problem of detecting unit roots in time series data is treated as a problem of multiple decisions instead of a testing problem, as is otherwise common in the econometric and statistical literature. The multiple decision design is based on a distinction between continuous primary and discrete secondary parameters. Four examples for such multiple decision designs are considered: first- and second-order integrated univariate processes; cointegration in a bivariate model; seasonal integration for semester data; seasonal integration for quarterly data. In all cases, restricted optimum decision rules are established based on Monte Carlo simulation.

Zusammenfassung: Die Bestimmung von Einheitswurzeln in Zeitreihendaten wird als multiples Entscheidungsproblem behandelt und nicht als Hypothesentest-Problem, wie es sonst in der ökonometrischen und statistischen Literatur üblich ist. Der verwendete entscheidungstheoretische Ansatz benützt eine Unterscheidung zwischen stetigen Primärparametern und diskreten Sekundärparametern. Vier Beispiele für die Anwendung des Ansatzes werden im Detail behandelt: univariate Prozesse mit unbekannter Integrationsordnung; Kointegration in bivariaten Modellen; saisonale Integration bei Halbjahresdaten; saisonale Integration bei Quartalsdaten. In allen Fällen werden optimale Entscheidungsregeln mittels Monte-Carlo-Simulation gefunden.

Keywords: Multiple decisions, unit roots, autoregressive processes

1 Introduction

Much of the recent literature on the analysis of macroeconomic time series focuses on the problem of making decisions on their degree of non-stationarity (for a good survey of the literature, see BANERJEE, DOLADO, GALBRAITH and HENDRY, 1993). Within this framework, researchers are particularly interested in whether the time series at hand has to be differenced once or twice or probably not at all to justify the assumption of covariance stationarity for the filtered series. Series requiring differencing once are usually called difference-stationary or first-order integrated. Additionally, in the joint analysis of two or more time series, researchers are interested in whether linear combinations of difference-stationary series may already be stationary. In this case, the linearly combined series are called cointegrated (for details, see the seminal paper by ENGLE and GRANGER, 1987).

Interest in cointegration has been instigated by technical problems as well as by economic theory. With regard to technical matters, it can be shown easily that differencing of cointegrated series leads to inefficient estimation due to a loss of information on low frequencies, even though individual series are difference-stationary. With regard to economic theory, evidence on long-run features is often interpreted as reflecting theoretical considerations on long-run equilibrium relations. Frequently quoted economic examples of this type are the long-run income elasticity of consumption, purchasing power parity, and the joint movement of interest rates with different terms to maturity.

Until the later 1980s, decisions on whether data sets suggest differencing, double differencing, or cointegration, were mainly based on the univariate testing procedure developed by FULLER (1976) and by DICKEY and FULLER (1979). Also in multivariate problems, these decisions tended to be based on a primary "cointegrating" regression and secondary residual analysis (see, e.g., ENGLE and GRANGER, 1987, and PHILLIPS and OULIARIS, 1990). JOHANSEN (1988) presented an efficient alternative framework for making such decisions. He suggested to determine the number of cointegrating relations by testing sequences and to proceed by conducting conditionally efficient estimation. PANTULA (1989) took up the idea of sequential testing for deciding upon the number of unit roots in univariate series.

In consequence, current integration/cointegration analysis is dominated by two main strands of statistical techniques. The first class of methods is characterized by easy handling and inefficiency caused by univariate residual analysis under limited information. The second class of methods relies on full system estimation but is faced with the usual problems of making decisions by sequential hypothesis testing. Alternatively, some researchers have used Bayesian methods, currently still with less impact on economic users. In the tradition of objectivist Bayesian statistics, most of them relied on continuous prior distributions designed to capture the researcher's lack of information before conducting the experiment. For a survey, see UHLIG (1994). An interesting case of a subjectively elicited mixed prior is given by KADANE, CHAN and WOLFSON (1996).

Here, a comprehensive framework for the problem of estimating the number of unit roots in univariate and multivariate situations is presented. In contrast to the bulk of the literature, this is not seen as a testing but as an estimation problem in the tradition of multiple decisions. In contrast to most Bayesian contributions to the literature but conforming with the Bayesian handling of model selection problems, a uniform prior is assumed on the decision parameters leading to mixtures of discrete and continuous distributions on the primary model parameters. Section 2 outlines the formal background. Section 3 presents examples and some evidence on corresponding decision bounds generated by Monte Carlo simulations. Section 4 concludes.

2 Estimating Discrete Parameters

Wherever possible, we would like to enhance the formal correspondence between the estimation of continuous and of discrete parameters. The reluctance by many researchers,

at least in the fields of applied economics, to call discrete parameter estimation problems by that name has probably led to occasional confusion. Often, discrete problems are called "model selection" or "sequential testing". The former expression deserves special attention. Following AKAIKE (1974), SCHWARZ (1978), and others, a sizable literature on discrete estimation problems has emerged. Most of the contributions focus on special applications, such as identifying the lag orders in ARMA models or subset selection in regressions. For two reasons, we will not adopt the label "model selection" here. Firstly, many researchers in applied fields link a special content to the task of model selection and do not see the formal equivalence to a discrete estimation problem. In practice, a model (or a class of models) may be selected on the basis of diagnostic statistics or of subject matter considerations and then this formal equivalence gets even more blurred. Secondly, a substantial part of the model selection literature itself does not focus on simply estimating discrete parameters. The now classical work of AKAIKE (1974) and the contribution by SAN MARTINI and SPEZZAFERRI (1984) are examples for this view where discrete parameter estimation is seen as an intermediate step in maximizing the utility evolving from a specific model choice. The recent contribution by PHILLIPS (1996) also falls into this category that is characterized by defining the distance between models via their forecasting performance and the Kullback-Leibler information. In contrast, we see discrete parameter estimation as the ultimate aim and, in consequence, define the distance between models solely via the discrete parameters of interest.

In concordance with, e.g., the work of HANNAN and DEISTLER (1988) or SCHWARZ (1978), we use the expression "estimation" for both continuous and discrete parametric approximation. In contrast, testing problems appear whenever one out of two hypotheses is given the preferred position of a "null hypothesis" and the researcher's loss is asymmetric because of subject matter considerations or of any reasons that permit a formal equivalence to quality control problems.

2.1 The Nested Problem

We consider the situation that observations are generated by an unknown member of a collection of distributions characterized by a parameter θ taken from a parameter space Θ. The parameter θ will also be called the *primary parameter*. Θ is the union of $p+1$ disjoint classes $\Theta_j, j = 0, ..., p$. Without including this in a formal definition, we may envisage class Θ_j as being characterized by a certain "feature" occurring j times. Then, $\bigcup_{i=0}^{j} \Theta_i$ represents the event of the feature occurring at most j times. In particular, Θ_0 represents the subset of Θ where the feature of interest is absent. The observer would like to make decisions on whether $\theta \in \Theta_j, j = 0, ..., p$. We will call the class index j the *secondary parameter* and $\Xi = \{0, ..., p\}$ the secondary parameter space. After estimating j, the user could also be interested in estimating θ within Θ_j. This problem, however, will be set aside within this study.

This point deserves attention as most advances in the fields of model selection appear to view estimation of the primary parameter as the final aim. In contrast, we focus exclusively on the classification problem, which will become obvious from the specification of

the loss function. Hence, the primary parameter may also be viewed as "nuisance" parameter in the decision problem, though we will restrict the notion of "nuisance parameter" to parts of the primary parameter.

In line with the above motivation concerning the problem of determining the frequency of occurrence of certain features, at first we restrict attention to the case that the model classes (or parameter sets) are ordered by an inclusion sequence

$$\Theta_0 \subset \bar{\Theta}_1, \Theta_1 \subset \bar{\Theta}_2, \ldots, \Theta_{p-1} \subset \bar{\Theta}_p \tag{1}$$

$\bar{\Theta}_i$ denotes the topological closure of Θ_i relative to the assumed topology in Θ. We will assume throughout that Θ is a topological space. Stronger assumptions will be used when necessary. Since in most interesting applications - including those considered in Section 3 - Θ is a subset of the m-dimensional Euclidean space $\mathbf{R}^m$, such assumptions are not unduly restrictive. In this case, we note that closure refers to the topology of Θ and not of the Euclidean space and, typically, neither Θ nor any Θ_i will be closed within the Euclidean space. We further note that, in this case, the trace of $(\mathbf{R}^m, d)$ naturally defines (Θ, d) as a metric space.

(1) implies that Θ_j is "small" relative to all Θ_i with $j > i$. Therefore, we will refer to (1) as the *nested problem.* To derive potential theoretical results and to exclude notorious cases, one may also impose the stricter condition

$$\bigcup_{k=0}^{j} \Theta_k = \bar{\Theta}_j \qquad j = 0, \ldots, p. \tag{2}$$

The choice function defined by

$$\kappa : \begin{cases} \Theta \to \Xi \\ \theta \mapsto \kappa(\theta) \quad \text{if} \quad \theta \in \Theta_{\kappa(\theta)} \end{cases} \tag{3}$$

maps the typically continuous primary parameter space onto the discrete secondary parameter space. The discrete secondary parameter $\kappa(\theta)$ summarizes all interesting information in θ, all other information is viewed as "nuisance" information. From now on, we will also use κ for the secondary parameter as we think it will not create confusion with the function $\kappa(.)$. The fact that, from the viewpoint of the decision problem, no penalty arises from selecting a member $\theta^* \neq \theta_0$ if the true primary parameter is θ_0, as long as $\kappa(\theta^*) = \kappa(\theta_0)$, must not be confounded with the fact that, within the sample, estimation of the secondary parameter can depend critically on the location of θ. To assess the precision achieved in estimating the secondary parameter, we need to define a distance measure [1] on Ξ. Concentrating on the Euclidean case, (Θ, d) and the function $\kappa(.)$ could be used to define a metric on Ξ which, however, is not very useful. The nested inclusion sequences (1) or (2) would imply that $\inf\{d(\theta_i, \theta_j) : \theta_i \in \Theta_i, \theta_j \in \Theta_j\} = 0$. Alternatively - remembering our initial interpretation of the nested problem - we adopt the logical position of viewing e.g. $j = 3$ to be "closer" to $j = 2$ than to $j = 1$ and we will expressly

[1] The expression "distance measure" is used as a free expression in contrast to a "metric" that must fulfill certain properties. Formally, a distance measure is just non-negative and symmetric.

use the squared distance measure

$$d_\kappa(i,j) = (i-j)^2. \tag{4}$$

We note that d_κ is the square of a metric on Ξ but it is not a metric as the triangle inequality does not hold. The researcher attempts to minimize the distance d_κ between his/her estimate of $\kappa(\theta)$ and the true value. We note that d_κ corresponds to measuring efficiency of continuous estimates via their variance.

Later we will define weighting priors on each of the classes Θ_i which could also be used to define a metric on Ξ via the average distance between points in two classes Θ_i and Θ_j, provided this weighted average distance is finite. We do not follow this route as we do not think that this average distance corresponds to the user's loss function.

Typically, the discrete secondary parameter is estimated indirectly by first estimating the continuous primary parameter θ. The estimate for θ is a random variable

$$\hat{\theta} : \begin{cases} (\Omega, \mathcal{A}, P) \to \Theta \\ x_n = (X_1, \ldots, X_n) \mapsto \hat{\theta}(x_n) \end{cases} \tag{5}$$

as the observations x_n are realizations of a random variable on the indicated probability space. The sample size is n. We assume that the estimator (5) is consistent. A good estimator of the primary parameter is certainly crucial for what follows. In most applications, the estimator is some approximation to the maximum likelihood estimator under the information that $\theta \in \Theta$. After making the decision on the secondary parameter, the observer may return to this problem and replace (5) by a more efficient estimator under the information that $\theta \in \Theta_j$ for fixed j.

A naive suggestion for constructing an estimator for the secondary parameter would be

$$\hat{\kappa}_N = \kappa(\hat{\theta}). \tag{6}$$

This is, however, unattractive because of the very definition of the nested problem (1) or (2). In finite samples, $\hat{\theta}$ usually - e.g., in all regression or time series problems that assume a continuous error distribution - has a continuous probability density, hence the topological smallness of $\Theta \setminus \Theta_p$ is reflected by the probability measure and $P(\hat{\theta} \in \Theta_p) = 1$. Thus, $P(\hat{\kappa}_N = p) = 1$. This property holds for every finite n, and the estimator is inconsistent.

The inclusion sequences (1) and particularly (2) have incited many researchers to solve the estimation problem via hypothesis testing. Hypothesis tests are constructed with the null hypothesis $H_0 : \theta \in \Theta_j$ and the alternative Θ_{j+1} or $\Theta \setminus \Theta_j$. An estimate for the secondary parameter is obtained by a certain stopping rule in such a testing sequence. There are four methods of this type in current usage:

(i) Test $\Theta_0 \cup \ldots \Theta_{p-1} = \bar{\Theta}_{p-1}$ against Θ_p; if rejected stop and $\hat{\kappa} = p$; otherwise test $\bar{\Theta}_{p-2}$ against Θ_{p-1}; if rejected stop and $\hat{\kappa} = p-1$; ... ; if no rejection $\hat{\kappa} = 0$.

(ii) As in (i) but always test $\bar{\Theta}_i$ against $\Theta \setminus \bar{\Theta}_i$.

(iii) Test Θ_0 against Θ_1; if accepted stop and $\hat{\kappa} = 0$; otherwise test $\bar{\Theta}_1$ against Θ_2; if accepted stop and $\hat{\kappa} = 1$; ... ; if everything is rejected $\hat{\kappa} = p$.

(iv) As in (iii) but always test $\bar{\Theta}_i$ against $\Theta \setminus \bar{\Theta}_i$.

The testing sequences (i) and (ii) correspond to the currently favored general-to-specific tests (see, e.g., YAP and REINSEL, 1995). (iii) and (iv) are specific-to-general. (iii) only works if rejection of Θ_j against Θ_{j+1} is guaranteed if Θ_{j+k} holds with $k \geq 2$. Asymptotically these properties are guaranteed by (2) and therefore all four testing sequences are consistent in the sense of test consistency.

Test consistency is defined by an asymptotic test power of unity for all parameter values belonging to the alternative but does not require the null to be accepted with probability one in large samples if it is true. If both properties are required, one speaks of "full (or complete) consistency". The testing steps with a fixed significance level are individually consistent but not fully consistent. Viewed as estimators for κ, they define inconsistent procedures unless $\kappa = p$. Reducing the significance level to zero asymptotically, one can construct consistent estimators of κ for many cases of empirical relevance. [2] For example, in the nested problem of autoregressive order selection, such a consistent estimation procedure was considered by POETSCHER (1983). The very popular estimators based on testing sequences with fixed significance level - usually 5% - will be summarily called *testing estimators*.

In finite samples, the properties of testing estimators differ across problems. Typically, (i) and (ii) yield a tendency toward small-sample upward biases and (iii) and (iv) toward downward biases. These tendencies can be counteracted by modifying significance levels.

The distance measure d_κ can be used to generate a different type of estimators. In analogy to e.g. least-squares estimation, let us assume that the investigator endeavors to minimize the loss function

$$l(\hat{\kappa}, x) = (\hat{\kappa}(x(\theta, \omega)) - \kappa(\theta))^2 = d_\kappa(\hat{\kappa}, \kappa). \tag{7}$$

The arguments are random variables and the right-hand side in (7) is unobserved. However, one could try to minimize expected loss given fixed θ:

$$E_\theta l(\hat{\kappa}, x) = \int_\Omega (\hat{\kappa}(x(\theta, \omega)) - \kappa(\theta))^2 dP_\theta(\omega) \tag{8}$$

In statistical decision theory, this function is called the risk function (see, e.g., FERGUSON, 1967). (8) is definitely not constant in κ and usually not constant in θ for given κ. Unlike in some classical problems, it is also not possible in general to solve the minimization problem analytically as this would require some knowledge about the small sample distribution of the primary parameter estimate. This turns out to be intractable in most applications. In order to make (8) operable in principle, one could try to finally define an estimator

$$\hat{\kappa} \quad \text{minimizes} \quad E E_\theta l(\hat{\kappa}, x) = \int_\Theta \int_\Omega (\hat{\kappa}(x(\theta, \omega)) - \theta)^2 dP_\theta(\omega) dQ(\theta). \tag{9}$$

[2] Many researchers are aware of this problem but deem it to be unimportant for the practitioner (see e.g. Theorem 12.7 by JOHANSEN, 1995b). Also, some Bayesians point out the complete consistency of their tests (see PHILLIPS and PLOBERGER, 1994) achieved by asymptotic reduction of significance levels.

However, (9) requires a definition of a probability measure Q on the parameter space Θ to define a weighting scheme. This will be done here. An alternative could be to minimize the supremum of the risk instead of a weighted average and would lead to the so-called minimax rules.

A formal Bayesian problem such as (9) naturally brings up the question of how to interpret the measure Q, as this is an issue where opinions vary considerably. We would like to avoid interpreting Q as a prior distribution reflecting prior beliefs about the parameter θ. It is simply used as an auxiliary weighting scheme for a decision problem. For such a decision problem among the discrete secondary parameter values, it appears logical to attribute the same weight to each of these values. In the Bayesian interpretation, this amounts to a non-informative prior over the secondary parameter space. In contrast to Bayesian analysis, we will not focus on posterior distributions but rather stick to the classical and supposedly more user-relevant problem of making discrete point decisions. For an example of the genuinely Bayesian point of view, see KADANE et al. (1996).

The uniform distribution on the secondary parameter space Ξ does not uniquely specify the distribution over the primary parameter Θ. To this aim, we define a weighting scheme on each $\Theta_j = \kappa^{-1}(\{j\})$ separately. In concordance with the uniform weighting for the secondary parameter, one may consider to define Q as uniform on Θ_j. This seems to be reasonable if Θ_j is bounded and convex. [3] However, if Θ_j is unbounded, the uniform law may not be properly defined. We would like to avoid the use of diffuse improperly defined priors as, in most practical applications, the weight given to unusual "far-away" parameter values is unacceptable.

In many applications, $\Theta = \Theta_{(1)} \times \Theta_{(2)}$ such that any parameter vector will consist of two parts $\theta = (\theta_1, \theta_2)$ where θ_1 is restricted to a bounded convex set and θ_2 is not restricted within a multidimensional Euclidean subspace. Cross restrictions are conceivable but the separation is important as it permits the construction of a uniform distribution on the subspace $\Theta_{(1)}$. Sometimes, this partition can be attained by a continuous transformation of the parameter space, starting from a given primary parameterization. Then, we consider the transformed parameterization as the "natural" one, assuming that classification of any θ into the Θ_j is independent of θ_2, i.e., $\kappa(\theta_1, \theta_2)$ is constant in θ_2. This convention does not define the parameterization uniquely, as e.g. any continuous one-to-one transformation of a given $\Theta_{(1)}$ onto itself defines an equally valid parameterization. We also do not require the dimension of $\Theta_{(1)}$ to be minimal. Typically, selection of a specific coordinate system is determined by the practitioner's concern rather than by formal properties. As an example, autoregressive models of fixed order have a convenient and natural parameterization if θ_1 consists of the coefficients and θ_2 of a possible mean. Decisions on the number of unit roots can be made based on θ_1, and $\Theta_{(1)}$ is bounded and convex. We could adopt the re-parameterization due to DICKEY and FULLER (1979) and thus minimize the dimension of $\Theta_{(1)}$ but this coordinate space is probably less natural. In contrast, the coefficient coordinates of vector autoregressions are not bounded within the Euclidean space, hence for decisions on cointegration a re-parameterization, as e.g. suggested by JOHANSEN (1988), is inevitable.

[3] Boundedness and convexity are not necessary but are used here to exclude atypical cases.

Now, assuming each $\Theta_j(1)$ to be bounded and convex, with $\Theta_j = \Theta_{j(1)} \times \Theta_{j(2)}$ and $\kappa(\theta_1, \theta_2)$ to be constant in θ_2, we can define Q as continuous uniform on $\Theta_{j(1)}$. This convention expresses the researcher's lack of information as well as the preference for the natural coordinate system. θ_2 is regarded as $nuisance$. Though the nuisance θ_2 is assumed not to influence the decision on the secondary parameter in population and in larger samples, we may permit a certain degree of dependence of the expected loss in (8) on θ_2 in finite samples. If θ_2 is defined on some higher-dimensional product of the real line, one could, e.g., impose standard normal distributions as weighting schemes for these nuisance parameters.

It is worth while to compare a so constructed distribution on Θ with prior distributions used in the literature. Firstly, uniform priors are widely avoided as they may produce strange results in some cases and are not invariant to transformations of the coordinates in Θ. This is less of a problem if a specific parameterization exists that is agreed upon as "natural" by most researchers. Secondly, mixed priors have been rarely used in the examples that will constitute our main focus of interest. For autoregressive processes, previous research has given positive weight to parts of the parameter space that are non-admissible a priori, such as explosive processes. The likely intention of this positive weighting of non-admissible parameter values is to draw attention to the admissible boundary a posteriori. In this interpretation, though the zero weighting of an interesting hypothesis and mixing of continuous and discrete distributions is avoided, it may be difficult to see the equivalence between the assumed "prior" and the researcher's true prior if such a one is hypothesized to exist and to be reasonable.

In the following examples, evaluation of optimum decision bounds will be based entirely on Monte Carlo simulation. A detailed description of the Monte Carlo technique is given in Section 3. The complicated metric imposed on the primary parameter space prevents analytical derivations, excepting the simplest case $\Xi = \{0,1\}$. Even numerically, however, (9) can hardly be solved directly for all possible estimators $\hat{\kappa}$. In restricting the considered class of decision rules in order to admit a numerical search for conditionally optimal solutions, one must focus on those rules where the loss relative to the unrestricted optimum is likely to be small. Under some regularity conditions - e.g. monotonous likelihood ratios - it can be shown by statistical theory that decision rules based on sufficient statistics and likelihood ratios are optimal in some sense. Not in all of our problems the corresponding criterion statistics are sufficient but it will always be assumed that the practitioner is primarily interested in keeping the decision rules simple. However, note that in the following, - in the notation of the nested problem - plausible restrictions such as $\Theta_j \subset \hat{\kappa}^{-1}(j), j = 0, \ldots, p-1$, are in general violated. We further note that basing decisions on likelihood-ratio type statistics facilitates the comparison to classical analysis in the framework of the testing estimator.

2.2 The Multiple Binary Problem

In the nested problem, the secondary parameter space Ξ appears to be naturally ordered. This corresponds well to cases where, for example, the number of non-zero or unit eigen-

values in a matrix are estimated. Ξ will always be equivalent to a finite sequence of natural numbers, such as $\{0,1,2,...,p\}$, or possibly the whole of $\mathbf{N_0}$. In the multiple binary problem, the elements of Ξ are k-tuples of binary numbers, such as (0,1,0,1). Formally, $\Xi = \{0,1\}^k$ for some k. This corresponds well to problems where k interesting and mutually (logically) independent features are either absent or present in the data. The set of decisions or secondary parameter values is also reminiscent of the power set over $0,...,k$ - note that this is not the σ-field used to construct a probability space but the set of elementary events - or of a Boolean algebra of order k. Therefore, we could also call it the *lattice problem.*

To handle the lattice problem in a similar way to the nested problem, we have to define a distance measure. Two extensions of the quadratic distance measure are conceivable. Firstly, one may use

$$d_1((a_1,\ldots,a_k),(b_1,\ldots,b_k)) = \sum_{i=1}^{k}(a_i - b_i)^2. \tag{10}$$

All the entries a_i and b_i are either 0 or 1, hence d_1 weights the maximum distance just by k. This corresponds to a linear weighting of large distances and does not appear to penalize large errors sufficiently. We will therefore use

$$d_k((a_1,\ldots,a_k),(b_1,\ldots,b_k)) = \left[\sum_{i=1}^{k}(a_i - b_i)^2\right]^2. \tag{11}$$

Again note that these distance measures are not metrics on Ξ as triangle inequalities fail to hold. Simple transforms would be metrics but would be uninteresting for our purposes. However, after forming expectations, metrics on probability spaces can be defined by taking e.g. square roots of the expectation of (4) or (10) or fourth roots of the expectation of (11).

We finally assume that the secondary parameter space Ξ is the whole of $\{0,1\}^k$ and that Q attributes the same weight to each k-tuple. In other words, the prior weighting will be uniform over all k-tuples.

Most observations can be transferred directly from our handling the nested problem to the multiple binary problem. Interestingly, even the classical treatment in the literature has been equivalent. Typically, one of the two cases - the "feature" being present or being absent - is reflected in a "generic" subset of the primary parameter space. The non-generic feature is then used as "null hypothesis" and is "tested" against the generic alternative. Obviously, a main problem concerns the assumptions about the features at different entries of the k-tuple. Routinely, classical testing chooses the convenient way of testing the non-generic feature at entry i under the (maintained) assumption that the non-generic feature also holds at all entries $j \neq i$. This design expresses a strong a priori belief in the non-generic features at all entries and can run into severe problems when more than one feature turns out to be generic. Alternatively, F-type and portmanteau tests assume the generic feature at all entries "under the alternative" and face the verdict of "low power". This low power is a consequence of the fact that the procedure classifies all mixed k-tuples such as (0,1,0,1) either as (0,0,0,0) or as (1,1,1,1).

As a viable alternative, we view the multiple binary problem as an estimation problem, where the secondary parameter is estimated such that the expected double squared loss expressed by the distance function d_k is minimized. Again, uniform weighting is assumed on the classes of primary parameters defined by $\Theta_a = \kappa^{-1}(a)$ and $a = (a_1, \ldots, a_k)$, or, in the common presence of unbounded nuisance parameters, uniform weighting on some $\Theta_{a(1)}$ in a convenient coordinate system.

Under the name of parameter subset selection, the lattice problem has been treated in the literature by BAUER et al. (1988). They prove that a consistent estimator for the secondary parameter exists if a form of uniform convergence holds for estimation in the primary parameter space. Such results ensure that an asymptotic risk of zero can be attained by decision rules that are t-tests or generalizations thereof. The problem remains how to minimize risk in finite samples. Whereas consistency, i.e. asymptotic zero risk, is independent of the distance function, the optimization of finite-sample performance may depend on the loss criterion.

A further generalization to "multiple nested problems" is straightforward. In this case, each feature j can appear with the frequency $l \in \{1, ..., p_j\}$ or not at all. In some applications, p_j will be constant over all j and the secondary parameter space will be equivalent to $0, ..., p^k$. Unsurprisingly, even this complicated discrete estimation problem has been routinely handled by sequences of binary tests with fixed significance levels in much of the literature leading to inconsistent secondary parameter estimates.

3 The Examples

The discrete parameter estimation technique outlined in Section 2 is applied to four simple problems of time series econometrics. In accordance with the minimization problem (9) and the loss functions (4) and (11) decision bounds are calculated as follows.

1. A primary parameter is randomly drawn from the prior distribution. This is done by first drawing the secondary parameter κ from a discrete uniform law and then the primary parameter θ according to the prior on Θ_κ.

2. The selected θ defines a data-generating process. A trajectory of n observations is generated from this process. All error processes in the construction are assumed as n.i.d.(0,1).

3. A small number of statistics is calculated from the trajectory and stored.

4. The first three steps are repeated n_s times. Each time the true secondary parameter and the relevant statistics are stored.

5. Tentative values for decision bounds on the statistics are selected and the approximate expected loss is evaluated.

6. The decision bounds are varied until a minimum of the expected loss is obtained. This minimization is conducted by a grid search involving several refining steps.

The value n_s defines the precision of the calculated bounds. In all examples, it was set to at least 10,000. The number of observations n was varied among several values that may conform to the size of typical economic data sets.

The so obtained minimizing bounds and the value of the expected loss at the minimum are then tabulated. FORTRAN codes of the procedures used for all examples can be obtained from the author on request.

3.1 Order of Integration in Second-order Autoregressions

We consider the second-order autoregressive model

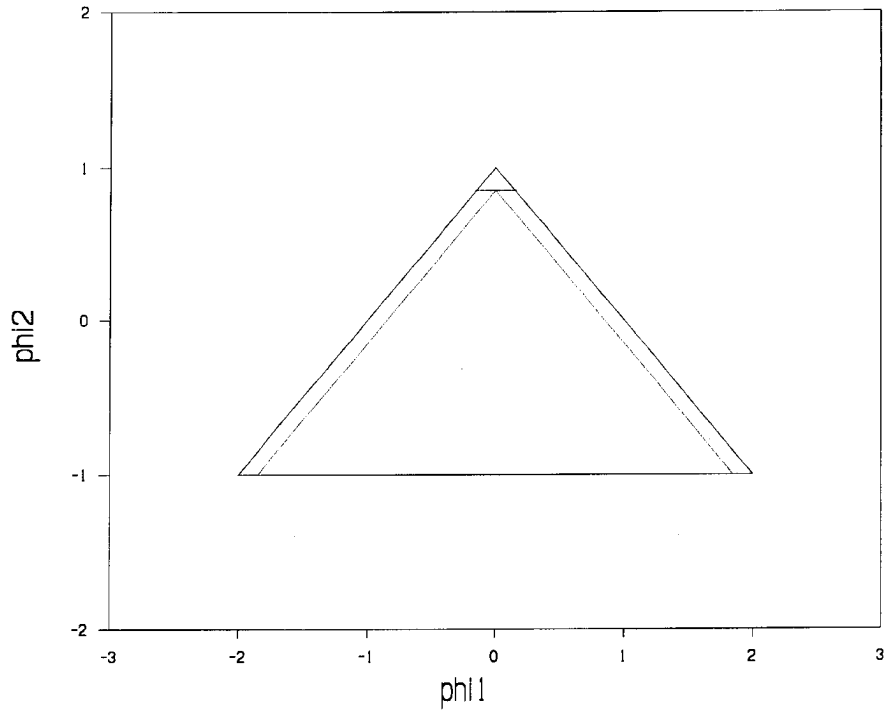
$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t \tag{12}$$

with n.i.d. $(0, \sigma^2)$ errors ϵ_t. It is well known that all sensible combinations of the parameters (ϕ_1, ϕ_2) are situated in and on a triangle flanked by the three lines

$$\phi_1 + \phi_2 = 1 \qquad -\phi_1 + \phi_2 = 1 \qquad \phi_2 = -1. \tag{13}$$

See our Figure 1 for a geometric interpretation.

Figure 1: The second-order difference equation. Stable solutions lie inside the triangle.



In the following, we will refer to this triangle as the SODE triangle for "second- order difference equations" whose stability conditions are reflected in it (e.g., see HAMILTON, 1994). All parameter combinations outside the triangle define anticipative or explosive

processes [4] and will therefore be excluded from the investigation. The set of sensible parameter values consists of the inner part

$$\Theta_2 = \{(\phi_1, \phi_2) \in \mathbf{R}^2 | \phi_1 + \phi_2 < 1, -\phi_1 + \phi_2 < 1, \phi_2 > -1\}$$

and the boundary of the triangle. All parameter values in Θ_2 define stationary AR(2) processes. The boundary of the triangle defines homogeneous non-stationary processes that are also called integrated processes. The maybe best known example $(\phi_1, \phi_2) = (1, 0)$ is the random walk. All points on the north-east boundary

$$\Theta_1 = \{(\phi_1, \phi_2) | \phi_1 + \phi_2 = 1, 0 < \phi_1 < 2\}$$

define first-order integrated processes. These are characterized by exactly one root of +1 in their characteristic polynomial and equivalently by the fact that they become stationary after one first-differencing transformation. The south-east corner point

$$\Theta_0 = \{(2, -1)\}$$

defines a second-order integrated process. It is the "double random walk"

$$X_t = \sum_{s=0}^{t} \sum_{r=0}^{s} \epsilon_r$$

and is the only process of its kind among the AR(2) processes. The other parts of the triangle boundary will be excluded for the moment. They are related to processes with very dominant periodicity, including the "mirror image" of the random walk $X_t = -X_{t-1} + \epsilon_t$. These will be examined more closely in Example 3. Hence, Θ is assumed to contain the interior of the SODE triangle and its north-east boundary half-closed by the south-east corner point. The remainder of the SODE boundary is not included in Θ.

Obviously, the design of this problem fulfills our assumptions (2) for a nested problem. The secondary parameter space Ξ is {0,1,2}. We note that $2 - \kappa$ is the order of integration of the process. In the literature, most authors have used the testing estimator based on approximate or exact ML estimates of the coefficients ϕ_1 and ϕ_2. 5% test boundaries were fixed by simulation or numerical integration as the asymptotic distribution of the LR statistic is a known transformation of Brownian motion integrals. As was already observed, the testing estimator with fixed significance level is inconsistent (see JOHANSEN, 1995a, and PANTULA, 1989).

To evaluate the asymptotic risk of the testing estimator, one may build on the following approximation. The exact asymptotic bias can be calculated from the formula given by JOHANSEN (1995a). If $\kappa = 2$, then the estimator is consistent and the asymptotic loss is 0. If $\kappa = 1$, there is a 5% chance of selecting $\kappa = 2$ and a 95% chance of uncovering the true value. Asymptotic loss is 0.05/3 because of the uniform prior weights assigned

[4] The difference equation $X_t = \phi X_{t-1} + \epsilon_t$ with $|\phi| > 1$ defines an explosive process if it is interpreted causally, with future evolving from the past. It defines a stationary process if it is interpreted anticipatively, with past evolving from the future. Neither one of the two cases corresponds to a useful description of economic data. Similar remarks hold for the second-order equation (12).

to the three values of the secondary parameter. If $\kappa = 0$, there is a probability of 0.05 of incorrect asymptotic "rejection", i.e., selection of different values of κ. Assuming the two "testing" steps to be approximately independent, given κ=0, the asymptotic loss becomes $(0.05 \cdot 0.95 + 4 \cdot 0.05^2)/3$, as the sequence of two incorrect "rejections" yields a loss of 4. The total asymptotic risk of the testing estimator is 0.03583... More efficient estimators have to be gauged against this number.

A consistent estimator with an asymptotic risk of 0 is the Bayes-rule estimator. In our case, its form would be:

$$\hat{\kappa} = \arg\max_j \int_{\Theta_j} f_\theta(x) d\theta$$

for $x = (x_1, \ldots, x_n) \in \mathbf{R}^n$ being the observed time series. This is a very simple case for applying the Bayes rule as each $\Theta_j = \kappa^{-1}(\{j\}), j = 0, 1, 2$, is completely expressed in the two parameters ϕ_1 and ϕ_2 and the assumed prior is uniform on Θ_j. The Bayes-rule estimator is consistent and minimizes the risk defined by the trivial distance function

$$d_B(i,j) = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j. \end{cases}$$

To attain a minimum for our more complicated quadratic distance measure d_κ defined by (4), we took refuge to Monte Carlo simulation. We note that d_κ expresses the view that a random walk is "closer" to a stationary process than the double unit root process $X_t = 2X_{t-1} - X_{t-2} + \epsilon_t$, which naturally extends our idea that 1 is closer to 0 than 2 is. This "intensity of incorrect decision-making" is reflected by the distance and the risk function.

Clearly, the requirement of asymptotic zero risk cannot define a decision rule uniquely. On the other hand, the theoretical optimum decision rules for a given finite sample size can be uncomfortably complex. In accordance with practitioners' needs, here simple and immediately operable decision rules will be preferred. A class of such simple decision rules is defined by the following design

1. select $\hat{\kappa} = 0$ if $\hat{\phi}_1 > 2 - b_1$
2. select $\hat{\kappa} = 1$ if $\hat{\phi}_1 + \hat{\phi}_2 > 1 - b_2$ and $\hat{\kappa} = 0$ is not selected
3. select $\hat{\kappa} = 2$ if neither $\hat{\kappa} = 0$ nor $\hat{\kappa} = 1$ is selected.

The optimum bounds b_1 and b_2 vary with the sample size and converge to 0 for large samples. It is easily seen that the so defined estimator $\hat{\kappa}$ is consistent if the coefficient estimators are. The estimator of the secondary parameter is consistent if the estimator of the primary parameters is consistent. Exact maximum likelihood, least squares, and the method-of-moments Yule-Walker rule all define consistent estimators of the primary parameters. In this Monte Carlo study, the least squares estimator is used as it is simple to calculate and therefore much in general use. However, it occasionally yields inadmissible

coefficient estimates that are treated just according to the general decision rule. The Yule-Walker estimate is unattractive in nearly non-stationary situations.

Our choice of bounds corresponds closely to the classical solution of the testing estimator. In fact, FULLER (1976) and also some later authors used the estimated coefficients proper, rather than using likelihood-ratio test statistics, for making decisions on whether unit roots are present. Of course, the binary decision problem between Θ_0 and Θ_1 is uninteresting as the Bayes rule defines an easy-to-use estimator that, in this case, also minimizes d_κ risk.

Table 1: Monte Carlo bounds for estimating the number of unit roots in a univariate AR(2) model. 10,000 replications were conducted.

n	b_1	b_2	risk
100	0.15	0.12	0.0564
200	0.08	0.08	0.0343
500	0.04	0.04	0.0145

Table 1 reports the results from our Monte Carlo simulation. For the smallest sample size n=100, the simulated bounds coincide well with the 5% bounds given in the literature. For larger sample sizes, their slower convergence toward 0 relative to the testing bounds becomes palpable. The bounds correspond to hypothesis tests with different size but nevertheless the achieved minimum risk may serve as a guideline in roughly suggesting that, in the absence of tables such as our Table 1, for n=500, decisions should be based on 2.5% rather than on 5% significance bounds. One may also compare the indicated decision bounds with the optimum achieved by the Bayes-rule estimator. All simulations were redone with the 0-1 loss function but the differences in optimum solutions were rather small so they are not reported.

3.2 Rank of Cointegrating Matrix

The first example can also be seen as estimating the rank of a certain matrix evolving from the state-space transition AR(1) form of the univariate AR(2) model

$$\begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix}$$

$$\vec{X}_t = \mathbf{T}\vec{X}_{t-1} + (\epsilon_t \quad 0)'.$$

If the state-space transition matrix $\mathbf{T}$ has all its eigenvalues smaller than 1, the autoregressive process is (asymptotically) stationary. If it has exactly one eigenvalue equal to 1, the process is first-order integrated. The process is second-order integrated if both eigenvalues of $\mathbf{T}$ are equal to unity. One could also consider the form

$$\Delta\vec{X}_t = (\mathbf{T} - \mathbf{I})\,\vec{X}_{t-1} + (\epsilon_t \quad 0)'.$$

Now, for $j = 1, 2$, Θ_j corresponds to the matrix **T**-**I** having rank j. The estimation problem of the secondary parameter becomes equivalent to *estimating the rank* of a stochastic matrix.

A similar problem evolves in truly multivariate time series analysis. Consider the so-called error-correcting representation of a bivariate first-order vector autoregression (see ENGLE and GRANGER, 1987)

$$\begin{bmatrix} \Delta X_t \\ \Delta Y_t \end{bmatrix} = (\mathbf{\Phi} - \mathbf{I}) \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}.$$

We assume that det(**I**-$\mathbf{\Phi}z$) has no roots ξ with $|\xi| < 1$ or $|\xi| = 1$ but $\xi \neq 1$. A further regularity condition excludes cases of second-order integration in one of the variables (see ENGLE and GRANGER and below). Here, the rank of the matrix $\mathbf{\Phi} - \mathbf{I}$ is particularly interesting. If it is 0, the processes X_t and Y_t are dynamically unrelated random walks. If it is 2, the bivariate process is stationary. If it is 1, both individual processes X and Y are first-order integrated but there is a stationary linear combination. Economists interpret this stationary cointegrating vector as expressing a long-run equilibrium relationship in the system. Obviously, this is again a nested problem in the sense of our definition.

In order to elicit a weighting measure on a parameter space Θ, the notion of a generic event is needed. Very informally, we define a generic event by being true on a subset of Θ that is so large as compared to Θ that any reasonable mass distribution is likely to assign a probability mass of 1 to the subset. We deliberately do not give the definition based on continuous mass distributions, as we would like to allow for discontinuities where they are plausible for the application. In our sense, a generic event is defined partly by its mathematical and partly by its subject matter properties. The opposite case, which is assigned a mass of 0 on the basis of similar considerations, will be called a non-generic event.

In this problem, primary parameters could be the elements of $\mathbf{\Phi}$ or $\mathbf{\Phi} - \mathbf{I}$. The classification of Θ into the three classes depends critically on the eigenvalues of $\mathbf{\Phi} - \mathbf{I}$. Admissible eigenvalues of $\mathbf{\Phi}$ lie in (-1,1), hence admissible eigenvalues of $\mathbf{\Phi} - \mathbf{I}$ are in (-2,0), with the borderline case 0 corresponding to unit roots. In this example, it is not so "natural" to choose a specific parameterization and therefore it is not so easy to find an appropriate weighting scheme. Here, the following idea was selected. The matrix $\mathbf{\Phi} - \mathbf{I}$ can be expressed via its Jordan canonical form:

$$\mathbf{\Phi} - \mathbf{I} = L \begin{bmatrix} \lambda_1 & \delta \\ 0 & \lambda_2 \end{bmatrix} L^{-1} \tag{14}$$

For "most" matrices, the element δ is 0. $\delta = 1$ for some matrices with $\lambda_1 = \lambda_2$. For the design of the following simulation study, we assume that this case plays little role. It is probably not very costly to exclude an event such as $\lambda_1 = \lambda_2$ anyway as it is non-generic. In particular, the case $\lambda_1 = \lambda_2 = \delta = 1$ is excluded which would entail a second-order integrated component process. A further difficulty is much more important. The Jordan canonical form is only valid in general if complex eigenvalues are admitted. For real matrices, these must be complex conjugates. The Jordan matrix can be represented in an

all-real form but we chose to exclude complex eigenvalues altogether. In Examples 3 and 4, this problem will be taken up again. Some cursory simulations allowing for complex conjugate eigenvalues proved that the results are not sensitive to our all-real design.

In this bivariate model, note that complex conjugates imply $|\lambda_1| = |\lambda_2|$, i.e., another apparently non-generic event. However, this kind of argument is probably faulty. The SODE triangle (Figure 1) shows a bottom area bordered by a dashed parabola corresponding to conjugate complex eigenvalues. This lower part covers two thirds of the entire area. Note, however, that it only touches upon Θ_1 at the corner point Θ_0. The discrimination of complex-rooted stationary processes from unit-root processes is probably a minor problem as compared to "average" real-rooted cases. In higher-dimensional models, this restriction may be more critical.

The matrix L in the Jordan decomposition can be any matrix provided it is non-singular. It is not uniquely determined. To reduce the effect of the non-uniqueness with respect to scaling, the innocuous normalization $l_{ii} = 1, i = 1, 2$, is imposed. The off-diagonal elements are allowed to take on any real values as long as these values do not succeed in making L singular, which again is a non-generic event but was not excluded a priori. The primary parameters $l_{ij}, i \neq j$, are treated as unbounded nuisance of the type θ_2 and are weighted according to a standard normal distribution. In later experiments, it may be interesting to vary this weighting on θ_2 and evaluate the sensitivity. We presume that the θ_2 weighting is unimportant.

In summary, we use uniform continuous prior weighting for (λ_1, λ_2) and standard normal priors for the l_{ij} with $i \neq j$ in (14) over the subspace Θ_2 of stationary processes. For the cointegrating processes Θ_1 and the "fully integrated" processes Θ_0, one or two of the eigenvalues λ_i are set at 0.

To check on the rank of $\mathbf{\Phi - I}$, a decision criterion could rely on the eigenvalues $\lambda_1, \lambda_2, |\lambda_1| \leq |\lambda_2|$, of $(\mathbf{\Phi - I})(\mathbf{\Phi - I})'$, as the number of non-zero eigenvalues of this symmetric matrix corresponds to the rank of $\mathbf{\Phi - I}$. We preferred to use squared canonical correlations between (X, Y) and $(\Delta X, \Delta Y)$, $\rho_1 \leq \rho_2$, instead, as suggested by JOHANSEN (1988). These are related to the likelihood ratio and can be extended easily to account for conditional influences in higher-order models or for correlation among $\epsilon_{1t}, \epsilon_{2t}$. It is shown easily that $\rho_j = 0$ if and only if the rank of $\mathbf{\Phi - I}$ is less than $3 - j$. Also, $\rho_j = 0$ if and only if $\lambda_j = 0$, provided that $\mathbf{\Phi - I}$ is diagonalizable. However, note that the prior weighting distribution was uniform on (-2,0) for λ_1, λ_2 but not on (0,1) for ρ_1, ρ_2.

Results from this bivariate cointegration experiment are summarized in Table 2. Actual decisions on the secondary parameter were based on sample estimates of the squared canonical correlations, in concordance with the likelihood-ratio analysis by JOHANSEN (1988). Our decision rule was defined in the following way:

1. Calculate the squared canonical roots and order them $0 \leq \hat{\rho}_1 \leq \hat{\rho}_2 \leq 1$.

2. Choose the stationary model if $\hat{\rho}_1 \geq b_1$.

3. Otherwise, choose the cointegrated model if $\hat{\rho}_2 \geq b_2$.

4. Otherwise, choose the fully integrated model.

This decision rule is similar in spirit to the classical eigenvalue test suggested by JOHANSEN (1988) as an alternative to the trace test whose fractiles are tabulated there. One may envisage the difference between the two decision rules - firstly, ours and Johansen's eigenvalue test and, secondly, Johansen's trace test - by plotting $\hat{\rho}_1$ and $\hat{\rho}_2$ in a plane where the permitted area is bounded by a triangle as the eigenvalues have been ordered by $\hat{\rho}_1 \leq \hat{\rho}_2$. The b_2 rule corresponds to a horizontal line whereas the trace test rule corresponds to a 45° negatively sloped line. Both "cut off" the area around the origin that indicates fully integrated processes. In both variants, the rule on the smaller eigenvalue corresponds to a vertical line. Table 2 allows a comparison between our procedure and the classical one under the caveat that the decision rules are slightly different with respect to ρ_2.

Table 2: Monte Carlo bounds for estimating the cointegrating rank in a bivariate AR(1) model. 10,000 replications were conducted. Approximate bounds suggested by a 5% significance level in the trace test for cointegration by JOHANSEN (1988) and the risk of this classical decision rule are given in square brackets.

n	b_1		b_2		risk	
100	0.045	[0.041]	0.127	[0.075]	0.0778	[0.0914]
200	0.026	[0.021]	0.070	[0.038]	0.0447	[0.0665]
300	0.022	[0.014]	0.053	[0.026]	0.0329	[0.0564]
400	0.019	[0.010]	0.039	[0.019]	0.0254	[0.0548]
500	0.015	[0.008]	0.038	[0.015]	0.0207	[0.0523]

It is not surprising that the risk of the classical decision rule substantially exceeds the optimum risk, as the classical test operates on an entirely different concept of risk that it tries to minimize. Its risk appears to settle down at values slightly above 5% at $n = 500$, which reflects its inconsistency. In contrast, our procedure attains 2% at $n = 500$. If $n = 100$, the usual 5% critical values roughly match those evolving from the multiple decision problem. In contrast, for $n = 500$, the significance level of the classical tests would have to be lowered to 1% to establish this equivalence. In consequence, for $n = 100$, b_1 corresponds roughly to Johansen's trace value whereas, for $n = 500$, b_1 is 1.8 times as large as the classical decision bound. On the other hand, b_2 is always much larger than the classical bound, which indicates that the classical procedure tends to avoid the fully integrated model even in small samples. In summary, substantially more fully integrated and some more co-integrated processes are found by the multiple decision procedure, these cases both gaining at the cost of stationary solutions. To put it conversely, the classical procedure appears to find uncomfortably many covariance-stationary processes when the true model is integrated. [5]

[5] Note that also PHILLIPS and PLOBERGER (1994) find more integrated models than previously used classical tests.

Note that the shape of the decision rule per se does not change much between the classical and our multiple decision framework. The main difference is in the significance levels not in the decision criterion.

3.3 Multiple Binary Problems: Seasonal Unit Roots

Let us take up the SODE triangle again. In Example 1, we had excluded the triangle boundary except for the north-east line segment, closed in the south-east by Θ_0 and open at the north corner. In particular since the publication of the article by HYLLEBERG et al. (1990, HEGY), econometricians have focused on cyclical and seasonal non-stationarity possibly explicable by unit roots at -1 alone or at both -1 and +1, extending the hitherto conducted approaches restricted to the unit root at +1. This model with "integration", i.e., spectral singularities, at the long-run and at the Nyqvist frequency seems particularly interesting for semester (half-yearly) data, whereas additional roots at the conjugate complex pair $\pm i$ may be considered for quarterly data (see Example 4). In Example 3, we concentrate on the semester case and on the root at -1.

Second-order autoregressive processes with exactly one unit root at -1 are found on the open north-west boundary line segment. The south-west corner point has second-order integration at -1. This case appears to be of mere academic interest and is unlikely to be found in economic reality. Hence, just like the explosive cases, the south-west corner point will be assigned zero weight. The north pole corresponds to integration both at +1 and at -1. This is the autoregressive process

$$X_t = X_{t-2} + \epsilon_t.$$

HYLLEBERG et al. (1990) and other authors have found that such processes provide reasonable descriptions of trending economic variables with substantial changes in their seasonal pattern. Hence, we do want to consider this case. In summary, we now have four subsets of the overall SODE triangle parameter space:

$\Theta_{\pm} = \{(0,1)\}$ ... integrated at long run and at Nyqvist frequency

$\Theta_{+} = \{(\phi_1,\phi_2) | \phi_1 \in (0,2), \phi_1 + \phi_2 = 1\}$ ... integrated at long run only

$\Theta_{-} = \{(\phi_1,\phi_2) | \phi_1 \in (-2,0), \phi_1 - \phi_2 = -1\}$ ... integrated at Nyqvist frequency only

$\Theta_S = \{(\phi_1,\phi_2) \in \mathbf{R}^2 | \phi_1 + \phi_2 < 1, -\phi_1 + \phi_2 < 1, \phi_2 > -1\}$ ... stationary

In the notation used in Section 2.2, the four events can also be coded in binary form as (1,1), (1,0), (0,1), (0,0), in this order, with the first entry corresponding to the unit root at 1 and the second entry to the unit root at -1.

The decision situation corresponds to the multiple binary or lattice problem introduced in Section 2.2. Note that there are two nested paths

$$\bar{\Theta}_{\pm} \subset \Theta_{+}, \bar{\Theta}_{+} \subset \Theta_S \quad \text{and} \quad \bar{\Theta}_{\pm} \subset \Theta_{-}, \bar{\Theta}_{-} \subset \Theta_S$$

The double-squared distance function introduced as (5) yields losses as outlined in the following table:

$d_{\pm}$	$\Theta_{\pm}$	Θ_{+}	Θ_{-}	Θ_{S}
$\Theta_{\pm}$	0	1	1	4
Θ_{+}	1	0	4	1
Θ_{-}	1	4	0	1
Θ_{S}	4	1	1	0

This table gives a cyclical definition of distance. A unit-root process of the long-run integrated type is supposed to be "closer" to a process integrated at both frequencies than to a process integrated at -1 only. This solution to the distance definition is probably debatable. In other multiple decision problems of similar type, this distance between the case of "exactly one object A" and "exactly one object B" obviously depends on the difference between objects A and B. When estimating the number of persons in a certain room or space, in most practical situations the distinction between men/women or black/white persons matters little and would not justify our cyclical design. On the other hand, the qualification of a good econometrician - who knows about economics and statistics as well - is probably closer to that of a statistician or of an economist than the two specialists' qualifications usually are among them. In our example, the properties of processes with roots at -1 and +1 are so strikingly different that the cyclical distance measure seems justified.

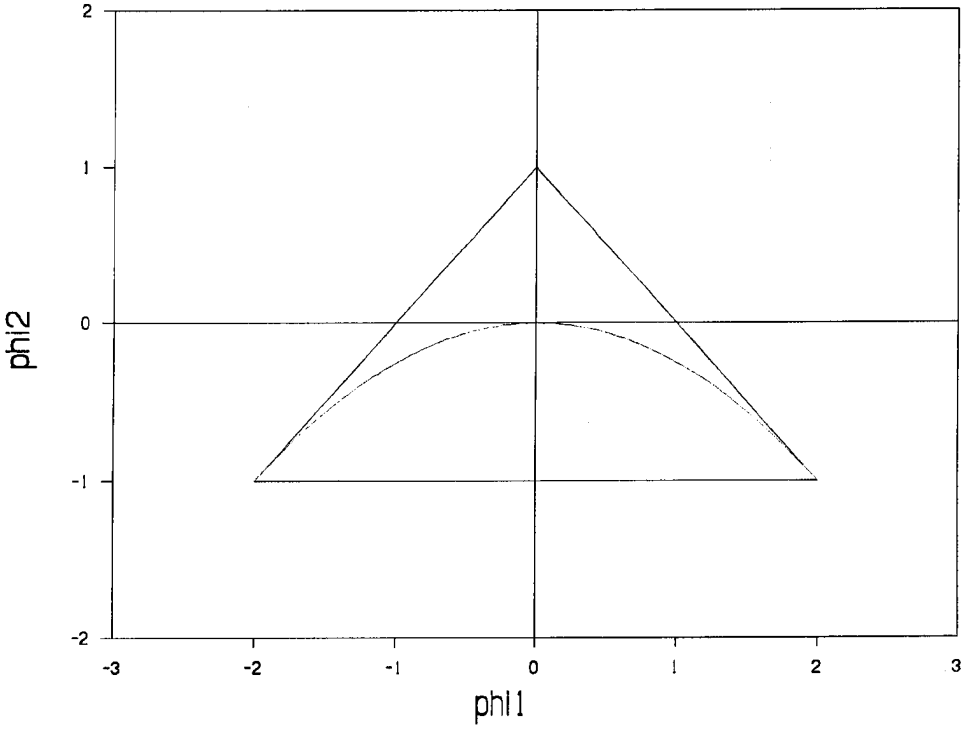
Monte Carlo simulations were conducted and an estimation of the secondary parameter in $\{(0,0),(0,1),(1,0),(1,1)\}$ was based on parallels to the north-east and north-west line segment and a horizontal beneath the north pole point. As the scheme appears to be perfectly symmetric between ϕ_1 and $-\phi_1$, there will be only two decision thresholds, b_1 describing the position of the horizontal and b_2 fixing the position of the skew parallels (see also Figure 2).

In summary, we have

$$\hat{\kappa} = \begin{cases} (0,0) \text{ if } \hat{\phi}_1 + \hat{\phi}_2 < 1 - b_2 \\ \qquad \text{and} \quad -\hat{\phi}_1 + \hat{\phi}_2 < 1 - b_2 \\ \qquad \text{and} \quad \hat{\phi}_2 < 1 - b_1 \\ (1,0) \text{ if } \hat{\phi}_1 + \hat{\phi}_2 > 1 - b_2 \\ \qquad \text{and} \quad \hat{\phi}_2 < 1 - b_1 \\ (0,1) \text{ if } -\hat{\phi}_1 + \hat{\phi}_2 > 1 - b_2 \\ \qquad \text{and} \quad \hat{\phi}_2 < 1 - b_1 \\ (1,1) \text{ otherwise.} \end{cases} \tag{15}$$

A technical problem derives from the fact that, as long as $b_1 < b_2$, the areas pointing to $\kappa = (0,1)$ and $\kappa = (1,0)$ may overlap. In this case, we select $\kappa = (1,0)$ whenever $\hat{\phi}_1 > 0$ and $\kappa = (0,1)$ otherwise. The results of some Monte Carlo simulations based on 50,000 replications are displayed as Table 3.

Figure 2: A sketch of the decision configuration



Since there are now four competing decisions, the risk is slightly higher than in Table 1, at corresponding sample sizes. Strikingly at odds with classical hypothesis test decisions, the optimum values for b_1 and b_2 are almost identical. It is interesting to have a closer look at these classical tests. The current recommendation seems to be to start by choosing among the secondary values $\kappa = (0, 1)$ or (1,1) and among $\kappa = (1, 0)$ or (1,1) separately. These tests with identical large-sample distributions correspond to our b_1 decision. If $\kappa = (1, 1)$ and $\kappa = (1, 0)$ are selected by the two separate tests, $\kappa = (0, 1)$ and $\kappa = (1, 1)$ are discarded and $\kappa = (1, 0)$ vs. $\kappa = (0, 0)$ are subjected to a third binary decision "due to the low power of the HEGY tests relative to the more specific DF test when

Table 3: Monte Carlo bounds for deciding among long-run, seasonal, and jointly long-run and seasonal non- stationarity in AR(2) models. 50,000 replications were conducted.

n	b_1	b_2	risk
100	0.133	0.136	0.0842
200	0.084	0.088	0.0488
300	0.060	0.066	0.0349
400	0.044	0.048	0.0288
500	0.042	0.046	0.0228

no seasonal unit root is present". The main conclusion to be drawn from suggesting this very complicated and hardly efficient procedure is that a priori confidence in the seasonal unit roots is lower than that in the more familiar cases $\kappa = (0, 1)$ and $\kappa = (0, 0)$.

Note that, for simplicity, the south-east corner point was excluded from consideration in Example 3. The union of Examples 1 and 3 can also be handled within our framework with $\Xi = \{0, 1, 2\} \times \{0, 1\}$ but it is not a lattice problem.

3.4 The so-called Univariate HEGY Model

The possibility of the joint presence of unit roots at different locations has been shown to complicate the handling in our multiple decision framework slightly but these difficulties can be overcome. In econometric practice, quarterly or monthly data are more common than semi-annual observations. For quarterly data, it is tempting to allow for the presence of homogeneous non-stationary influences deriving from the main frequencies $0, \pi/2, \pi$ though other frequencies would be conceivable. In econometrics, the main reference to this problem is again HEGY (1990). There, fourth-order autoregressive structures were considered. These were transformed into the form

$$\Delta_4 X_t = c_1 S(B) X_{t-1} + c_2 A(B) X_{t-1} + (c_3, c_4)(\Delta_2 X_{t-1}, \Delta_2 X_{t-2})' + \epsilon_t$$

with $S(B) = 1+B+B^2+B^3$, $A(B) = 1-B+B^2+B^3$, $\Delta_i = 1-B^i$, and B denoting the backshift or lag operator defined by $BX_t = X_{t-1}$. HEGY (1990) then suggest to conduct t- and F-type tests on the significance of the coefficients in order to find out about the potential significance of rejecting unit roots at 1 (the coefficient c_1), at -1 (the coefficient c_2), and at $\pm i$ (c_3 and c_4 jointly). As was already stated above, we want to develop alternatives to this classical framework which is, moreover, based on the assumption of "just local tests", with the remaining unit roots assumed as being present anyway. Only asymptotically, such cross-effects among effects at different seasonal frequencies vanish.

To handle the HEGY problem in our framework, we again rely on the double-squared distance measure for lattice problems (11). The secondary parameter can be coded as a 3-vector of binary numbers (a_1, a_2, a_3), convening that $a_1 = 1$ stands for the presence of a unit root at +1, $a_2 = 1$ for a unit root at -1, and $a_3 = 1$ for the complex pair $\pm i$. Then, e.g. (0,0,0) corresponds to the event of "no unit roots", and (0,1,1) to "no unit root at 1 but one each at -1 and $\pm i$". In detail, the distance measure is defined by

$$d_k\left(\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}\right) = \left(\sum_{j=1}^{3} |a_j - b_j|\right)^2 .$$

We note that all a_j and b_j elements are either 0 or 1 and that the maximum distance of 9 is, e.g., obtained between (1,0,0) and (0,1,1), i.e., between a process of the random-walk type and one that wholly consists of persistent seasonal cycles.

Next, the weighting distributions over the primary parameters within the classes have to be fixed. This is trivial for (1,1,1), since there is only one fourth-order process with all three unit roots present:

(a) (1,1,1) is simulated as $\Delta_4 X_t = \epsilon_t$.

(b) For (0,1,1), the process must look like $(1-\phi B)(1+B)(1+B^2)X_t = \epsilon_t$. We assume a uniform weighting prior on ϕ within the interval (-1,1). Similar rectangular priors can be chosen for (1,0,1).

(c) (1,1,0) is simulated using a uniform weighting prior over the SODE triangle to generate processes of type $(1 - B^2)(1 - \phi_1 B - \phi_2 B^2)X_t = \epsilon_t$.

(d) For (0,0,1), we use the design $(1 + B^2)(1 - \phi_1 B - \phi_2 B^2)X_t = \epsilon_t$ again over the SODE triangle for (ϕ_1, ϕ_2).

For (1,0,0) and (0,1,0), a counterpart to the SODE triangle in the three-dimensional space would be required. However, the structure of the stationarity area for the third-order difference equation is already quite involved. It is convex but not a simplex and does not have planes at all boundaries. We decided to use "brute force" instead for any order larger than two. For three lags, noting that the coefficients in a third-order stationary difference equation have maximum absolute values of (1,3,3,1), single coefficients were drawn from uniform random variables of (-3,3), (-3,3), and (-1,1), respectively.

(e) (1,0,0) and (0,1,0) are generated from $(1 - B)(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)X_t = \epsilon_t$ and $(1+B)(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)X_t = \epsilon_t$. The primary parameters (ϕ_1, ϕ_2, ϕ_3) are generated by draws from three uniform distributions. Stability of the difference equation is checked and (ϕ_1, ϕ_2, ϕ_3) are re-drawn if explosive roots have been found.

(f) (0,0,0) is generated from the full fourth-order design $(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \phi_4 B^4)X_t = \epsilon_t$. Maxima for $(|\phi_1|, |\phi_2|, |\phi_3|, |\phi_4|)$ are (4,6,4,1). Stability is checked and independent re-drawing is performed if necessary.

Across the classes, a uniform prior was assumed and hence each class obtains the relative weight of 0.125. Table 4 gives the results from a Monte Carlo experiment according to the outlined design. For each of the sample sizes 100 and 200, 80,000 replications were simulated. This gives approximately 10,000 replications for each specific model class. Table 4 does not only show the simulated bounds but also gives the matrix of correct and incorrect decisions in the experiment. Larger sample sizes probably are not relevant in practice, due to the fact that quarterly data are rarely available for time spans of more than 50 years, maybe excepting meteorological series.

4 Summary and Conclusion

Many problems of multiple decisions are usually handled by binary sequential testing decisions with much emphasis on keeping a "correct" constant size of the test components. The quality control framework may not correspond to the interest of the practitioner who

Table 4: Empirical frequencies of selecting the respective events of seasonal integration if the loss function is double quadratic. Number of replications is 80,000.

(a) $n = 100$

true	selected model							
	(0,0,0)	(1,0,0)	(0,1,0)	(1,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,1)
(0,0,0)	6824	1284	1424	97	314	21	22	0
(1,0,0)	89	8951	38	752	2	165	0	1
(0,1,0)	91	33	8939	769	3	0	181	4
(1,1,0)	8	444	436	9064	1	3	5	49
(0,0,1)	15	1	8	1	8868	543	563	18
(1,0,1)	1	83	0	4	207	9430	23	236
(0,1,1)	2	0	71	4	239	18	9399	254
(1,1,1)	2	8	13	156	76	757	742	8244

Bounds: b_1=0.060 b_2=0.062 b_3=0.126 Loss at minimum = 0.1444

(b) $n = 200$

true	selected model							
	(0,0,0)	(1,0,0)	(0,1,0)	(1,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,1)
(0,0,0)	7661	1124	983	50	152	10	6	0
(1,0,0)	40	9321	14	540	1	79	0	3
(0,1,0)	47	18	9277	601	0	0	77	0
(1,1,0)	2	270	222	9499	0	1	0	16
(0,0,1)	5	0	0	0	9194	429	378	11
(1,0,1)	0	28	0	1	77	9688	5	185
(0,1,1)	0	0	20	1	111	7	9667	181
(1,1,1)	0	3	1	65	10	450	355	9114

Bounds: b_1=0.045 b_2=0.041 b_3=0.082 Loss at minimum = 0.0876

intends to classify the data at hand into one out of a small number of categories. Two frequent forms of such problems have been considered, the nested and the lattice problem.

In the nested problem, the researcher is interested in estimating a naturally ordered discrete parameter. A related example of this type would be estimating the lag order of an autoregressive structure but this problem has been treated extensively in the literature (see also HANNAN and DEISTLER, 1988, Ch. 5). Information criteria have been shown to provide consistent estimates of the lag order and have widely replaced the less adequate sequential tests that lead to inconsistent estimates if significance levels are fixed. For the problem of estimating the order of integration in time series, a satisfactory treatment is still needed and our Examples 1 and 2 have contributed to that aim.

In the lattice problem, the researcher is interested in a number of features that could be

present in the data or not. A common example would be the inclusion/exclusion decision on possible regressor variables in linear regressions that is usually handled by t-tests and F-tests. However, in that case, many researchers may find themselves in a quality control situation and their decision may closely correspond to rejecting or accepting a subject matter theory. In contrast, in estimating seasonal unit root models, this view is less likely. The researcher rather attempts to find the one model out of 4 (Example 3) or 8 (Example 4) structures that most closely tracks the data at hand. We have provided a new and promising framework for making such decisions.

Much work remains to be done in the future. Amalgams of the nested and lattice models appear when a variety of features can be absent/present in varying numbers and the number associated to each feature is interesting. Such a situation evolves, e.g., in seasonal cointegration. Another situation obtains when the absence or presence of deterministic features - such as constants, trends, fixed cycles - is investigated jointly with the unit roots. Depending on the interest in the features per se, the presence or absence of the features may define distinct decision classes or may be treated as nuisance.

To find an optimum decision rule, we assumed squared loss for the secondary parameters which are the only parameters of interest here. Squared loss is a common concept in estimating continuous parameters and we feel that multiple decisions should be treated in a joint framework. Risk typically depends on all primary parameters and we adopted uniform weighting of these parameters over "natural" parameterizations, conscious of the fact that uniform weighting is not invariant to re-parameterizations. We also gave equal weights to each class (secondary parameter) considered. We finally insisted on the typical researcher's aim to make binary (not quantitative) decisions on the secondary parameters, leaving Bayesian grounds with the latter viewpoint.

Viewed from a Bayesian perspective, we stressed the importance of mixed (continuous-discrete) priors in typical situations of multiple decisions. In contrast, the continuous priors used by most Bayesian researchers in unit root estimation entail two severe problems. Firstly, they only achieve posterior mass for the non-generic classes by putting prior mass on non-admissible extensions of the parameter space, such as explosive processes. Secondly, they put probably undue emphasis on the problem of estimating primary parameters in a way that the researcher may not be interested if he/she is faced with the problem of making strictly binary decisions such as choosing among conflicting viewpoints arising from economic theory.

The four examples contain an interesting aspect that may seem unusual to Bayesian statisticians. The tabulated decision bounds in our framework may serve purposes similar to the usual tables of fractiles. However, most practitioners may be reluctant to use such tables unless they can be convinced that the most general structure as represented by Θ may be general enough to capture a model that could likely have generated the observed data. The answer to this point is a delicate matter, hence we will conclude with three possible ones. Firstly, we may say that the statistician defines the model. If it is a second-order autoregression, such tables can be used in accordance with the statistician's "window to reality". Secondly, if one suspects that the model is not general enough, one may repeat the numerical experiment with randomized nuisance, thus "opening the win-

dow". Thirdly, if further discrete parameters, such as the order of an autoregression, are of genuine interest, one may construct more classes and "enlarge" the window. Unfortunately, computing time may limit the dimension of the decision set in this latter solution.

Acknowledgments

The author wishes to thank all those who have commented on earlier versions of this paper, in particular Manfred Deistler, Benedikt Poetscher, Erhard Reschenhofer, more generally participants in seminars in Vienna, Prague, and St. Oswald, and last but not least two anonymous referees. The usual proviso applies.

References

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

BANERJEE, A., DOLADO, J., GALBRAITH, J.W., HENDRY, D.F. (1993). *Co-integration, Error-correction, and the Econometric Analysis of Non-Stationary Data.* Oxford University Press.

BAUER, P., B.M. POETSCHER, HACKL, P. (1988). Model Selection by Multiple Test Procedures. *Statistics* **19**, 39–44.

DICKEY, D.A., FULLER, W.A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association* **74**, 427–431.

ENGLE, R.F., GRANGER, C.W.J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica* **55**, 251–276.

FERGUSON, T.S. (1967). *Mathematical Statistics - A Decision Theoretic Approach.* Academic Press.

FULLER, W.A. (1976). *Introduction to Statistical Time Series.* Wiley.

HAMILTON, J.D. (1994). *Time Series Analysis.* Princeton University Press.

HANNAN, E.J., DEISTLER, M. (1988). *The Statistical Theory of Linear Systems.* Wiley.

HYLLEBERG, S., ENGLE, R.F., GRANGER, C.W.J., YOO, B.S. (1990). Seasonal Integration and Cointegration. *Journal of Econometrics* **44**, 215–238.

JOHANSEN, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* **12**, 231–254.

JOHANSEN, S. (1995a). A Statistical Analysis of Cointegration for I(2) Variables. *Econometric Theory* **11**, 25–59.

JOHANSEN, S. (1995b). *Likelihood-Based Inference in Cointegrated Vector Auto-regressive Models.* Oxford University Press.

KADANE, J.B., CHAN, N.H., WOLFSON, L.J. (1996). Priors for unit root models. *Journal of Econometrics* **75**, 99–111.

PANTULA, S.G. (1989). Testing for Unit Roots in Time Series Data. *Econometric Theory* **5**, 265–271.

PHILLIPS, P.C.B. (1996). Econometric Model Determination. *Econometrica* **64**, 763–812.

PHILLIPS, P.C.B., OULIARIS, S. (1990). Asymptotic Properties of Residual Based Tests for Cointegration. *Econometrica* **58**, 165-194.

PHILLIPS, P.C.B., PLOBERGER, W. (1994). Posterior Odds Testing for a Unit Root with Data-Base Model Selection. *Econometric Theory* **10**, 774–808.

SAN MARTINI, A., SPEZZAFERRI, F. (1984). A Predictive Model Selection Criterion. *Journal of the Royal Statistical Society* **46**, 296–303.

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

UHLIG, H. (1994). What Macroeconomists Should Know About Unit Roots - A Bayesian Perspective. *Econometric Theory* **10**, 645–671.

YAP, S.F., REINSEL, G.C. (1995). Estimation and Testing for Unit Roots in a Partially Nonstationary Vector Autoregressive Moving Average Model. *Journal of the American Statistical Association* **90**, 253–267.

Author's address:

Univ.-Doz. Dipl.-Ing. Dr. Robert M. Kunst
Institute for Advanced Studies
Stumpergasse 56, A-1060 Wien
E-Mail: kunst@ihssv.wsr.ac.at