Article

Cointegration and Unit Root Tests: A Fully Bayesian Approach

Marcio A. Diniz , Carlos A. B. Pereira and Julio M. Stern

Special Issue
Data Science: Measuring Uncertainties

Edited by
Prof. Dr. Carlos Alberto De Bragança Pereira , Prof. Dr. Adriano Polpo and Prof. Dr. Agatha Rodrigues

Several times series present deterministic or stochastic trends, which imply that the effects of these trends on the level of the series are permanent. Consequently, the mean and variance of the series will not be constant and will not revert to a long-term value. This feature reflects the fact that the stochastic processes generating these series are not (weakly) stationary, imposing problems to perform inductive inference using the most traditional estimators or predictors. This is so because the usual properties of these procedures will not be valid under such conditions.

Therefore, when modeling non-stationary time series, one should be able to properly detrend the used series, either by directly modeling the trend by deterministic functions, or by transforming the series to remove stochastic trends. To determine which strategy is the suitable solution, several statistical tests were developed since the 1970s by the frequentist school of statistical inference.

The Augmented Dickey–Fuller (ADF) test is one of the most popular tests used to assess if a time series has a stochastic trend or, for series described by auto-regressive models, if they have a unit root. When one is searching for long term relationships between multiple series under analysis, it is crucial to know if there are stationary linear combinations of these series, i.e., if the series are cointegrated. Cointegration tests were developed, also by the frequentist school, in the late 1980s [1] and early 1990s [2]. Only in the late 1980s did the Bayesian approach to test the presence of unit roots start to be developed.

Both unit root and cointegration tests may be considered tests on precise or sharp hypotheses, i.e., those in which the dimension of the parameter space under the tested hypothesis is smaller than the dimension of the unrestricted parameter space. Testing sharp hypotheses poses major difficulties for either the frequentist or Bayesian paradigms, such as the need to eliminate nuisance parameters.

The main goal of this article is to briefly review the shortcomings of the tests proposed by the Bayesian school and how they can be overcome by the Full Bayesian Significance Test (FBST). More specifically, we will compare its performance with the most used frequentist alternatives, the ADF

for unit roots, and the maximum eigenvalue test for cointegration. Since this is a review article, it is important to remark that the results presented here were published elsewhere by the same authors, see [3,4].

To accomplish this objective, we will define the FBST in the next section, also showing how it can be implemented in a general context. The following section discusses the problems of testing the existence of unit roots in univariate time series and how the Bayesian tests approach the problem. Section 4 then shows how the FBST is applied to test if a time series has unit roots and illustrates this with applications on a real data set. In the sequel, we discuss the Bayesian alternatives to cointegration tests and then apply the FBST to test for cointegration using real data sets. We conclude with some remarks and possible extensions for future work.

1. FBST

The Full Bayesian Significance Test was proposed in [5] mainly to deal with sharp hypotheses. The procedure has several properties, see [6,7], most interestingly the fact that it is only based on posterior densities, thus avoiding the necessity of complications such as the elimination of nuisance parameters or the adoption of priors with positive probabilities attached to sets of zero Lebesgue measure.

We shall consider general statistical models in which the parameter space is denoted by $\Theta \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$. A sharp hypothesis H assumes that θ, the parameter vector of the chosen statistical model, belongs to a sub-manifold Θ_H of smaller dimensions than Θ. This implies, for continuous parameter spaces, that the subset Θ_H has null Lebesgue measure whenever H is sharp. The sample space, the set of all possible values of the observable random variables (or vectors), is here denoted by $\mathcal{X}$.

Following the Bayesian paradigm, let $h(\cdot)$ be a probability prior density over Θ, $\mathbf{x} \in \mathcal{X}$, the observed sample (scalar or vector), and $L(\cdot \mid \mathbf{x})$ the likelihood derived from data $\mathbf{x}$. To evaluate the Bayesian evidence based on the FBST, the sole relevant entity is the posterior probability density for θ given $\mathbf{x}$,

$$g(\theta \mid \mathbf{x}) \propto h(\theta) \cdot L(\theta \mid \mathbf{x}).$$

It is important to highlight that the procedure may be used when the parameter space is discrete. However, when the posterior probability distribution over Θ is absolutely continuous, the FBST appears as a more suitable alternative to significance hypothesis testing. For notational simplicity, we will denote Θ_H by H in the sequel.

Let $r(\theta)$ be a reference density on Θ such that the function $s(\theta) = g(\theta \mid \mathbf{x})/r(\theta)$ is a *relative surprise*, (see [8], pp. 145–146) function. The reference density is important because it guarantees that the FBST is invariant to reparametrizations, even when $r(\theta)$ is improper, see [6,9]. Thus, when considering $r(\theta)$ proportional to a constant, the surprise function will be, in practical terms, equivalent to the posterior distribution. For the applications considered in this article, we will use the improper uniform density as reference density on Θ. The authors of [10] remark that it is possible to generalize the procedure using other reference densities such as neutral, invariant, maximum-entropy or non-informative priors, if they are available and desirable.

Definition 1 (Tangent set). *Considering a sharp hypothesis $H : \theta \in \Theta_H$, the tangential set of the hypothesis given the sample is given by*

$$\mathbb{T}_{\mathbf{x}} = \{\theta \in \Theta : s(\theta) > s^*\}. \tag{1}$$

where $s^* = \sup_{\theta \in H} s(\theta)$.

Notice that the tangent set $\mathbb{T}_{\mathbf{x}}$ is the highest relative surprise set, that is, the set of points of the parameter space with higher relative surprise than any point in H, being *tangential* to H in this sense. This approach takes into consideration the statistical model in which the hypothesis is defined, using several components of the model to define an evidential measure favoring the hypothesis.

Definition 2 (Evidence). *The Bayesian evidence value against H, $\overline{ev}$, is defined as*

$$\overline{ev} = P\left(\theta \in \mathbb{T}_{\mathbf{x}} \mid \mathbf{x}\right) = \int_{\mathbb{T}_{\mathbf{x}}} dG_{\mathbf{x}}(\theta), \qquad (2)$$

where $G_{\mathbf{x}}(\theta)$ denotes the posterior distribution function of θ and the above integral is of the Riemann–Stieltjes type.

Definition 2 sets $\overline{ev}$ as the posterior probability of the tangent set that is interpreted as an evidence value against H. Hence, the evidence value supporting H is the complement of $\overline{ev}$, namely, $ev = 1 - \overline{ev}$. Notwithstanding, ev is not evidence against $A : \theta \notin \Theta_H$, the alternative hypothesis (which is not sharp anyway). Equivalently, $\overline{ev}$ is not evidence in favor of A, although it is against H.

Definition 3 (Test). *The FBST is the procedure that rejects H whenever $ev = 1 - \overline{ev}$ is smaller than a critical level, ev_c.*

Thus, we are left with the problem of deciding the critical level ev_c for each particular application. We briefly discuss this and other practical issues in the following subsection.

1.1. Practical Implementation: Critical Values and Numerical Computation

Since ev (also called e-value) is a statistic, it has a sampling distribution derived from the adopted statistical model and in principle this distribution could be used to find a threshold value. If the likelihood and the posterior distribution satisfy certain regularity conditions. See [11], p. 436. [12] proved that, asymptotically, there is a relationship between ev and the p-values obtained from the frequentist likelihood ratio procedure used to test the same hypotheses. This fact provides a way to find, at least asymptotically, a critical value to ev to reject the hypothesis being tested.

In a recent review [7], the authors discuss different ways to provide a threshold for ev. Among these alternatives, we highlight the standardized e-value, which follows, asymptotically, the uniform distribution on $(0, 1)$. See also [13] for more on the standardized version of ev.

One could also try to define the FBST as a Bayes test derived from a particular loss function and the respective minimization of the posterior expected loss. Following this strategy, [10] showed that there are loss functions which result in ev as a Bayes estimator of $\phi = \mathbb{I}_H(\theta)$, where $\mathbb{I}_A(x)$ denotes the indicator function, being equal to one if $x \in A$ and zero otherwise, $x \notin A$. Hence, the FBST is in fact a Bayes procedure in the formal sense as defined by Wald in [14].

Table 1. Pseudocode to implement the FBST.

General algorithm: compute ev supporting hypothesis $H : \theta \in \Theta_H$
1. Specify the statistical model (likelihood) and prior distribution on Θ.
2. Specify the reference density, $r(\theta)$, and derive the relative surprise function, $s(\theta)$.
3. Find s^*, the maximum value of $s(\theta)$ under the constraint $\theta \in H$.
4. Integrate the posterior distribution on the tangent set—Equation (2)—to find $\overline{ev}$.
5. Find $ev = 1 - \overline{ev}$.

To compute the evidence value supporting H defined in the last section, we need to follow the steps showed in Table 1. Appendix A provides detailed information about the computational resources and codes used to implement the FBST in the examples presented in this work. After defining the statistical model and prior, it is simple to find the surprise function, $s(\theta)$. In step 3, one should find the point of the parameter space in H that maximizes $s(\theta)$, that is, to solve a problem of constrained numerical maximization. In several applications, this step does not present a closed form solution, requiring the use of numerical optimizers.

Step 4 involves the integration of the posterior distribution on a subset of Θ, the tangent set $\mathbb{T}_{\mathbf{x}}$ that can be highly complex. Once more, since in many cases it is fairly difficult to find an explicit

expression for $\mathbb{T}_{\mathbf{x}}$, one may use various numerical techniques to compute the integral. If it is possible to generate random samples from the posterior distribution, Monte Carlo integration provides an estimate of *ev*, as we will show in this work. Another alternative is to use approximation techniques, such as those proposed in [15], based on a Laplace approximation. We discuss how to implement such approximations for unit root and cointegration tests in [3,4].

2. Bayesian Unit Root Tests

Before presenting the Bayesian procedures used to test the presence of unit roots, let us fix notation. We will denote by y_t the *t*-th value of a univariate time series observed in $t = 1, \dots, T + p$ dates, where T and p are positive integers. The usual approach is to assume that the series under analysis is described by an auto-regressive process with p lags, $AR(p)$, meaning that the data generating process is fully described by a stochastic difference equation of order p, possibly with an intercept or drift and a deterministic linear trend, i.e.,

$$y_t = \mu + \delta \cdot t + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t \tag{3}$$

with ε_t i.i.d. $N(0, \sigma^2)$ for $t = 1, \dots, T + p$. Using the lag or backshift operator B, we denote y_{t-k} by $B^k y_t$, allowing us to rewrite (3) as

$$(1 - \phi_1 B - \dots \phi_p B^p) y_t = \mu + \delta \cdot t + \varepsilon_t \tag{4}$$

where $\phi(B) = (1 - \phi_1 B - \dots \phi_p B^p)$ is the autoregressive polynomial. The difference Equation (3) will be stable, implying that the process generating $\{y_t\}_{t=1}^{T+p}$ is (weakly) stationary, whenever the roots of the characteristic polynomial $\phi(z)$, $z \in \mathbb{C}$, lie outside the unit circle, since there may be complex roots. The set of polynomial operators, such as lag polynomials like $\phi(B)$, induces an algebra that is isomorphic to the algebra of polynomials in real or complex variables, see [16].

If some of the roots lie exactly on the unit circle, it is said that the process has unit roots. In order to test such a hypothesis statistically, (3) is rewritten as

$$\Delta y_t = \mu + \delta \cdot t + \Gamma_0 \, y_{t-1} + \Gamma_1 \Delta y_{t-1} + \dots + \Gamma_{p-1} \Delta y_{t-p+1} + \varepsilon_t \tag{5}$$

where $\Delta y_t = y_t - y_{t-1}$, $\Gamma_0 = \phi_1 + \dots + \phi_p - 1$ and $\Gamma_i = -\sum_{j=i+1}^{p} \phi_j$, for $i = 1, \dots, p-1$. If the generating process has only one unit root, one root of the complex polynomial $\phi(z)$,

$$1 - \phi_1 z - \phi_2 z^2 \dots \phi_p z^p,$$

is equal to one, meaning that

$$1 - \phi_1 - \phi_2 - \dots - \phi_p = 0$$

i.e., $\phi(1) = 0$, and all the other roots are on or outside the unit circle. In this case, $\Gamma_0 = 0$, the hypothesis that will be tested when modeling (5). Even though tests based on these assumptions verify if the process has a single unit root, there are generalizations based on the same principles that test the existence of multiple unit roots, see [17].

The search for Bayesian unit root tests began in the late 1980s. As far as we know, [18,19] were the first works to propose a Bayesian approach for unit root tests. The frequentist critics of these articles received a proper answer in [20,21], generating a fruitful debate that produced a long list of papers in the literature of Bayesian time series. A good summary of the debate and the Bayesian papers that resulted from it is presented in [22]. We will present here only the most relevant strategies proposed by the Bayesian school to test for unit roots.

Let $\theta = (\rho, \psi)$ be the parameters vector, in which $\rho = \sum_{i=1}^{p} \phi_i$ and $\psi = (\mu, \delta, \Gamma_1, \dots, \Gamma_{p-1})$. Assuming σ^2 fixed, the prior density for θ can be factorized as

$$h(\theta) = h_0(\rho) \cdot h_1(\psi \mid \rho).$$

The marginal likelihood for ρ, denoted by L_m, is:

$$L_m(\rho \mid \mathbf{y}) \propto \int_{\Psi} L(\theta \mid \mathbf{y}) \cdot h_1(\psi \mid \rho) \, d\psi.$$

where $\mathbf{y} = \{y_t\}_{t=1}^{T+p}$ is the observations vector, $L(\theta|\mathbf{y})$ the full likelihood, and Ψ the support of the random vector ψ. This marginal likelihood, associated with a prior for ρ, is the main ingredient used by standard Bayesian procedures to test the existence of unit roots. Even though the procedure varies among authors according to some specific aspects, mentioned below, basically all of them use Bayes factors and posterior probabilities.

One important issue is the specification of the null hypothesis: some authors, starting from [23], consider $H_0 : \rho = 1$ against $H_1 : \rho < 1$. Starting from [24], this is the way the frequentist school addresses the problem, but following this approach no explosive value for ρ is considered. The decision theoretic Bayesian approach solved the problem using the posterior probabilities ratio or Bayes factor:

$$B_{01} = \frac{L_m(\rho = 1 \mid \mathbf{y})}{\int_0^1 L_m(\rho \mid \mathbf{y}) \cdot h_0(\rho) \, d\rho}.$$

Advocates of this solution argue that one of the advantages of this approach is that the null and the alternative hypotheses are given equal weight. However, the expression above is not defined if $h_0(\rho)$ is not a proper density since the denominator of the Bayes factor is equal to the predictive density, defined just if $h_0(\rho)$ is a proper density. There are also problems if $L_m(\rho = 1|\mathbf{y})$ is zero or infinite.

The problem is approached by [20,25] by testing $H_0 : \rho \geq 1$ against $H_1 : \rho < 1$, considering explicitly explosive values for ρ. The main advantage of this strategy is the possibility to compute posterior probabilities like

$$P(\rho > 1 \mid \mathbf{y}) = \int_1^{\infty} g_m(\rho \mid \mathbf{y}) \, d\rho$$

defined even for improper priors on ρ, where g_m is the marginal posterior for ρ.

In [26], the authors do not choose ρ as the parameter of interest, examining instead the largest absolute value of the roots of the characteristic polynomial and then verifying if it is smaller or larger than one. Usually, this value is slightly smaller than ρ, but the authors argue that this small difference may be important. When this approach is used, unit roots are found less frequently. For an AR(3) model with a constant and deterministic trend, [26] derives the posterior density for the dominant root for the 14 series used in [27] and concluded the following: for eleven of the series, the dominant root was smaller than one, that is to say, the series were trend-stationary. These results were based on flat priors for the autoregressive parameters and the deterministic trend coefficient.

Another controversy is about the prior over ρ: [20] argues that the difference between the results given by the frequentist and Bayesian inferences is due to the fact that the flat prior proposed in [18] overweights the stationary region of ρ. Hence, he derived a Jeffreys prior for the AR(1) model: this prior quickly diverges as ρ increases and becomes larger than one. The obtained posterior led to the same results of [27], which will be discussed in detail in the following section. The critics of the approach adopted by Phillips in [20] judged the Jeffreys prior as unrealistic, from a subjective point of view. See the comments on Phillips's paper on the *Journal of Applied Econometrics*, volume 6, number 4, 1991. The subsequent papers of the same number support the Bayesian approach. This is a nonsensical objection if one considers that the Jeffreys prior is crucial to ensure an invariant inferential procedure, and invariance is a highly desirable property, for either objective or subjective reasons. See [28] for more on invariance in physics and statistical models.

A final controversial point concerns the modeling of initial observations. If the likelihood explicitly models the initial observed values (it is an *exact* likelihood), the process is implicitly considered stationary. In fact, when it is known that the process is stationary, and it is believed that the data

generating process is working for a long period, it is reasonable to assume that the parameters of the model determine the marginal distribution of the initial observations. In the simplest AR(1) model, this would imply that $y_1 \sim N(0, \sigma^2/(1-\rho^2))$. In this scenario, to perform the inference conditional on the first observation would discard relevant information. On the other hand, there is no marginal distribution defined for y_1 if the generating process is not stationary. Then, it is valid to use a likelihood conditional on initial observations. For the models presented here, we always work with the conditional likelihood. As argued in [18], inferences for stationary models are little affected by using conditional likelihoods, especially for large samples. He compares these inferences with the ones based on exact likelihoods under explicit modeling for initial observations.

3. Implementing the FBST for Unit Root Testing

We will now describe how to use the FBST to test for the presence of unit roots referring to the general model (5). It is also possible to include $q \in \mathbb{N}$ moving average terms in (3) to model the process, a case that will not be covered in this article but that, in principle, shall not imply major problems for the FBST.

$$\Delta y_t = \mu + \delta \cdot t + \Gamma_0\, y_{t-1} + \Gamma_1 \Delta y_{t-1} + \ldots + \Gamma_{p-1} \Delta y_{t-p+1} + \varepsilon_t, \tag{5}$$

where $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$ for $t = 1, \ldots, T + p$, recalling also that the hypothesis being tested is $\Gamma_0 = 0$. We slightly change the notation of the last section now using ψ to denote the vector $(\mu, \delta, \Gamma_0, \ldots, \Gamma_{p-1})$ and setting $\theta = (\psi, \sigma)$.

Recalling the steps to implement the FBST displayed in Table 1, we have just specified the statistical model. The likelihood, conditional on the first p observations, derived from the Gaussian model is

$$L(\theta \mid \mathbf{y}) = (2\pi)^{-T/2}\sigma^{-T}\exp\left\{ -\frac{1}{2\sigma^2} \cdot \sum_{t=p+1}^{T+p} \varepsilon_t^2 \right\}, \tag{6}$$

in which $\varepsilon_t = \Delta y_t - \mu - \delta \cdot t - \Gamma_0 y_{t-1} - \Gamma_1 \Delta y_{t-1} - \ldots - \Gamma_{p-1} \Delta y_{t-p+1}$. To complete step 1 of Table 1, we need a prior distribution for θ. For all the series modeled in this article, we will use the following non informative prior:

$$h(\theta) = h(\psi, \sigma) \propto 1/\sigma. \tag{7}$$

We are aware of the problems caused by improper priors applied to this problem when one uses alternative approaches, like those mentioned by [22]. However, one of our goals is to show how the FBST can be implemented even for a potentially problematic prior like this one. To write the posterior, we use the following notation:

$$\Delta Y = \begin{bmatrix} \Delta y_{p+1} \\ \Delta y_{p+2} \\ \vdots \\ \Delta y_{T+p} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & p+1 & y_p & \Delta y_p & \ldots & \Delta y_2 \\ 1 & p+2 & y_{p+1} & \Delta y_{p+1} & \ldots & \Delta y_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & T+p & y_{T+p-1} & \Delta y_{T+p-1} & \ldots & \Delta y_{T+1} \end{bmatrix}, \quad \psi = \begin{bmatrix} \mu \\ \delta \\ \Gamma_0 \\ \vdots \\ \Gamma_{p-1} \end{bmatrix},$$

being ΔY of dimension $T \times 1$, X of dimension $T \times (p+2)$ and ψ, $(p+2) \times 1$. Thanks to this notation, we can write, using primes to denote transposed matrices:

$$\sum_{t=p+1}^{T+p} \varepsilon_t^2 = (\Delta Y - X\psi)'(\Delta Y - X\psi) = (\Delta Y - \widehat{\Delta Y})'(\Delta Y - \widehat{\Delta Y}) + (\psi - \widehat{\psi})'X'X(\psi - \widehat{\psi}),$$

where $\widehat{\psi} = (X'X)^{-1}X' \cdot \Delta Y$ is the ordinary least squares (OLS) estimator of ψ and $\widehat{\Delta Y} = X\widehat{\psi}$ its prediction for ΔY. Thus, the full posterior is

$$g(\theta \mid \mathbf{y}) \propto \sigma^{-(T+1)} \exp\left\{ -\frac{1}{2\sigma^2}[(\Delta Y - \widehat{\Delta Y})'(\Delta Y - \widehat{\Delta Y}) + (\psi - \widehat{\psi})'X'X(\psi - \widehat{\psi})] \right\}, \tag{8}$$

a Normal-Inverse Gamma density.

Step 2 demands a reference density in order to define the relative surprise function. Since we will use the improper density $r(\theta) \propto 1$, the surprise function will be equivalent to the posterior distribution in our applications. Given this, to find s^* (Step 3), we need to find the maximum value of the posterior under the hypothesis being tested, in our case, $\Gamma_0 = 0$.

This maximization step is fairly simple to implement given the modeling choices made here: Gaussian likelihood, non informative prior and reference density proportional to a constant. The restricted (assuming H) posterior distribution is

$$g_r(\theta_r \mid \mathbf{y}) \propto \sigma^{-(T+1)} \exp\left\{ -\frac{1}{2\sigma^2}[(\Delta Y - \widehat{\Delta Y}_r)'(\Delta Y - \widehat{\Delta Y}_r) + (\psi_r - \widehat{\psi}_r)'X_r'X_r(\psi_r - \widehat{\psi}_r)] \right\}, \tag{9}$$

in which $\theta_r = (\psi_r, \sigma)$, ψ_r being vector ψ without Γ_0,

$$X_r = \begin{bmatrix} 1 & p+1 & \Delta y_p & \dots & \Delta y_2 \\ 1 & p+2 & \Delta y_{p+1} & \dots & \Delta y_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & T+p & \Delta y_{T+p-1} & \dots & \Delta y_{T+1} \end{bmatrix}, \quad \widehat{\psi}_r = (X_r'X_r)^{-1}X_r' \cdot \Delta Y, \text{ and } \widehat{\Delta Y}_r = X_r\widehat{\psi}_r,$$

that is, X_r is simply matrix X above without its third column, since under $H : \Gamma_0 = 0$ and the coefficient of the third column of X is Γ_0—see Equation (5)—$\widehat{\psi}_r$ is a least squares estimator of ψ_r and $\widehat{\Delta Y}_r$ denotes the predicted values for ΔY given by the restricted model. From (9), it is easy to show that the maximum a posteriori (MAP) estimator of θ_r is given by $(\widehat{\psi}_r, \widehat{\sigma}_r)$, with

$$\widehat{\sigma}_r = \sqrt{\frac{(\Delta Y - \widehat{\Delta Y}_r)'(\Delta Y - \widehat{\Delta Y}_r)}{T+1}}.$$

Plugging the values of $\widehat{\psi}_r$ and $\widehat{\sigma}_r$ into (9), we find s^*, as requested in Step 3. Step 4 will also be easy to implement thanks to structure of the models assumed in this section. Since the full posterior, (8), is a Normal-Inverse Gamma density, a simple Gibbs sampler allows us to obtain a random sample from such distribution, suggesting a Monte Carlo approach to compute $\overline{ev}$. From (8), the conditional posteriors of ψ and σ are, respectively,

$$g_\psi(\psi \mid \sigma, \mathbf{y}) \propto N(\widehat{\psi}, \sigma^2(X'X)^{-1}) \tag{10}$$

$$g_{\sigma^2}(\sigma^2 \mid \psi, \mathbf{y}) \propto IG\left(\frac{T+1}{2}, H\right) \tag{11}$$

in which $H = 0.5[(\Delta Y - \widehat{\Delta Y})'(\Delta Y - \widehat{\Delta Y}) + (\psi - \widehat{\psi})'X'X(\psi - \widehat{\psi})]$, IG denotes the Inverse-Gamma distribution and $\widehat{\psi}$ is the OLS estimator of ψ, as above. Appendix B brings the parametrization and the probability density function of the Inverse-Gamma distribution. With a sizable random sample from the full posterior, we estimate $\overline{ev}$ as the proportion of sampled vectors that generate a value for the posterior greater than s^*, found in Step 3. Hence, in Step 5, we only compute one minus the estimate of $\overline{ev}$ found in Step 4. The whole procedure is summarized in Table 2. For the implementations in this article we sampled 51,000 vectors from (8) and discarded the first 1,000 as a burn-in sample.

Table 2. Pseudocode to implement the FBST to unit root tests.

General algorithm: compute *ev* supporting hypothesis $H : \Gamma_0 = 0$ in model (5)
1. Statistical model: Gaussian; prior: $h(\theta) \propto 1/\sigma$.
2. Reference density: $r(\theta) \propto 1$; relative surprise function: $g(\theta \mid \mathbf{y})$.
3. Find s^*: (9) evaluated at $\widehat{\psi}_r$ and $\widehat{\sigma}_r$.
4. Gibbs sampler (from Equations (10) and (11)) to obtain N random samples of parameter vectors from (8). Evaluate the posterior at the sampled vectors and estimate $\overline{ev}$ as the proportion of N in which the evaluated values are larger than s^*.
5. Find $ev = 1 - \overline{ev}$.

Results

We implemented the FBST as described above to test the presence of unit roots in 14 U.S. macroeconomic time series, all with annual frequency, first mentioned in [27]. We used the extended series, analyzed in [23]. Appendix A brings more information on the data set and the computational resources and codes used to obtain the results displayed in Table 3 below.

Table 3 reports the names of the tested series, the number of available observations or sample size, the adopted value for *p*—as denoted in Equation (8)—if a linear (deterministic) trend was included in the model or not, the ADF test statistic and its respective *p*-value. We have used the computer package described in [29] to find the ADF *p*-values, available in the R library urca. The last two columns bring the posterior probability of non-stationarity, $\Gamma_0 \geq 0$, and the FBST e-values for the specified models. In order to obtain comparable results, we have adopted the models chosen by [22] for all the series. All the models considered the intercept or constant term, μ in (8).

The results show that the non-stationary posterior probabilities are quite distant from the ADF *p*-values. These results were highlighted in [18,19]. Considering the simplest AR(1) model, they argued that, once frequentist inference is based on the distribution of $\widehat{\rho}|\rho = 1$, the non-stationary posterior probabilities provide counterintuitive conclusions since the referred distribution is skewed. Their main argument is that Bayesian inference uses a distribution (marginal posterior of ρ) that is not skewed.

As mentioned before, ref. [20] claims that the difference in results between frequentist and Bayesian approaches is due to the flat prior that puts much weight on the stationary region. He proposed the use of Jeffreys priors, which restored the conclusions drawn by the frequentist test. Phillips argued that the flat prior was, actually, informative when used in time series models like those for unit root tests. Using simulations, he shows that " *[the use of a] flat prior has a tendency to bias the posterior towards stationarity. ... even when [the estimate] is close to unity, there may still be a non negligible downward bias in the [flat] posterior probabilities*". Notwithstanding, the e-values reported in the last column are quite close to the ADF *p*-values even using the flat prior criticized by Phillips.

Table 3. Unit root tests for the extended Nelson and Plosser data set.

Series	Sample Size	*p*	Trend	ADF	*p*-Value	$P(\Gamma_0 \geq 0\|\mathbf{y})$	e-Value
Real GNP	80	2	yes	−3.52	0.044	0.0005	0.040
Nominal GNP	80	2	yes	−2.06	0.559	0.0238	0.523
Real GNP per capita	80	2	yes	−3.59	0.037	0.0004	0.034
Industrial prod.	129	2	yes	−3.62	0.032	0.0003	0.028
Employment	99	2	yes	−3.47	0.048	0.0004	0.043
Unemployment rate	99	4	no	−4.04	0.019	0.0001	0.020
GNP deflator	100	2	yes	−1.62	0.778	0.0584	0.762
Consumer prices	129	4	yes	−1.22	0.902	0.1154	0.983
Nominal wages	89	2	yes	−2.40	0.377	0.0106	0.341
Real wages	89	2	yes	−1.71	0.739	0.0475	0.715
Money stock	100	2	yes	−2.91	0.164	0.0029	0.147
Velocity	119	2	yes	−1.62	0.779	0.0620	0.777
Bond yield	89	4	no	−1.35	0.602	0.0962	0.936
Stock prices	118	2	yes	−2.44	0.357	0.0103	0.349

4. Bayesian Cointegration Tests

Before starting our brief review of the most relevant Bayesian cointegration tests, we fix notation and present the definitions to which we will refer in the sequel.

All the tests mentioned here are based on the following multivariate framework. Let $\mathbf{Y}_t = [y_{1t} \ldots y_{nt}]'$ be a vector with $n \in \mathbb{N}$ time series, all of them assumed to be integrated of order $d \in \mathbb{N}$, i.e., have d unit roots. The series are said to be cointegrated if there is a nontrivial linear combination of them that has $b \in \mathbb{N}$ unit roots, $b < d$. We will assume that, as in most applications, $d = 1$ and $b = 0$, meaning that, if the time series in $\mathbf{Y}_t$ is cointegrated, there is a linear combination $\mathbf{a}'\mathbf{Y}_t$ that is stationary, where $\mathbf{a} \in \mathbb{R}^n$ is the cointegrating vector. Since the linear combination $\mathbf{a}'\mathbf{Y}_t$ is often motivated by problems found in economics, it is called a long-run equilibrium relationship. The explanation is that non-stationary time series that are related by a long-run relationship cannot drift too far from the equilibrium because economic forces will act to restore the relationship.

Notice also that: (i) the cointegrating vector is not uniquely determined since, for any scalar s, $(s \cdot \mathbf{a})$ is a cointegrating vector; and (ii) if $\mathbf{Y}_t$ has more than two series, it is possible that there is more than one cointegrating vector generating a stationary linear combination.

It is assumed that the data generating process of $\mathbf{Y}_t$ is described by the following vector autoregression with $p \in \mathbb{N}$ lags, denoted VAR(p), and given by:

$$\mathbf{Y}_t = \mathbf{c} + \Phi_0\mathbf{D}_t + \Phi_1\mathbf{Y}_{t-1} + \ldots + \Phi_p\mathbf{Y}_{t-p} + \mathbf{E}_t, \tag{12}$$

in which $\mathbf{c}$ is a $(n \times 1)$ vector of constants, $\mathbf{D}_t$ a vector $(n \times 1)$ with some deterministic variable, such as deterministic trends or seasonal dummies, Φ_i are $(n \times n)$ coefficients matrices and $\mathbf{E}_t$ is a $(n \times 1)$ stochastic vector with multivariate normal distribution with null expected value and covariance matrix Ω, denoted $N_n(\mathbf{0}, \Omega)$. This dynamic model is assumed valid for $t = 1, \ldots, T + p$, the available span of observations of $\mathbf{Y}_t$. As in the univariate case, one may include moving average terms in (12), i.e., lags for $\mathbf{E}_t$, but this, in principle, would not cause major problems in the Bayesian framework. Model (12) can be rewritten using the lag or backshift operator as

$$(I_n - \Phi_1 B - \ldots - \Phi_p B^p)\mathbf{Y}_t = \mathbf{c} + \Phi_0\mathbf{D}_t + \mathbf{E}_t, \tag{13}$$

where $\Phi(B) = I_n - \Phi_1 B - \ldots - \Phi_p B^p$ is the (multivariate) autoregressive polynomial and I_n denotes the n-dimensional identity matrix. The associate characteristic polynomial in this context will be the determinant of $\Phi(z)$, $z \in \mathbb{C}$. If all the roots of the characteristic polynomial lie outside the unit circle, it is possible to show that $\mathbf{Y}_t$ has a stationary representation—see [30]—such as Equation (12). In order to determine if this is the case, model (12) is rewritten as an (vectorial) error correction model (VECM):

$$\Delta\mathbf{Y}_t = \mathbf{c} + \Phi_0\mathbf{D}_t + \Gamma_1\Delta\mathbf{Y}_{t-1} + \ldots + \Gamma_{p-1}\Delta\mathbf{Y}_{t-p+1} + \Pi\mathbf{Y}_{t-1} + \mathbf{E}_t, \tag{14}$$

where $\Delta\mathbf{Y}_t = [\Delta y_{1t} \ldots \Delta y_{nt}]'$, $\Gamma_i = -(\Phi_{i+1} + \ldots \Phi_p)$ for $i = 1, 2, \ldots, p-1$ and $\Pi = -\Phi(1) = -(I_n - \Phi_1 - \ldots - \Phi_p)$. It is possible to show that, when all the roots of $\det(\Phi(z))$ are outside the unit circle, matrix Π in (14) has full rank, i.e., all the n eigenvalues of Π are n non null. If the rank of Π is null, this matrix cannot be distinguished from a null matrix, implying that the series in $\mathbf{Y}_t$ has at least one unit root and a valid representation is a VAR of order $p - 1$, i.e., model (14) without the term $\Pi\mathbf{Y}_{t-1}$. It is possible that the series in $\mathbf{Y}_t$ has two unit roots each, implying that the correct VECM must be written with $\Delta^2\mathbf{Y}_t$ as a dependent variable.

Finally, if the $(n \times n)$ matrix Π has rank r, $0 < r < n$, it has $n - r$ non null eigenvalues, implying that the series in $\mathbf{Y}_t$ has at least one unit root and its valid representation is given by the VECM in Equation (14). In this case, $\Pi = \alpha\beta'$, where α and β are matrices $(n \times r)$ of rank r. Matrix β denotes the one with the cointegrating vectors and matrix α is called the loading matrix, since it contains the weights of the equilibrium relationships. The tests developed in [2] focus on the rank of matrix Π.

The pioneer Bayesian works to study VAR models and reduced rank regressions are [31–33]. However, the main concern of these papers is to estimate the model parameters and their (marginal) posterior distributions. The usual approach is to assume a given rank for the long run matrix Π, and proceed with all the computations conditional on the given rank. The Bayesian initiatives to test the rank of the referred matrix are recent, the main reference for Bayesian inference on VECM's being [34].

To justify inferential procedures based on prespecified ranks of matrix Π, [22] argued that an empirical cointegration analysis should be based on economic theory, which proposes models obeying equilibrium relationships. According to this view, cointegration research should be "confirmatory" rather than "exploratory". Even though the advocated conditional inference is of simple implementation and very useful for small samples, [22] recognized that tests for the rank of matrix Π should be developed. To our knowledge, few initiatives with this purpose were developed up to now.

One common approach to test sharp hypotheses in the Bayesian framework is by means of Bayes factors. Testing the rank of matrix Π by Bayes factors implies several computational complications and requires the use of proper priors, as shown in [35]. Following an informal approach, [33] obtained the posterior distribution of the ordered eigenvalues of the "squared" long run matrix, $\Pi' \cdot \Pi$, obtained from a VAR model without assuming the existence of cointegration relations. As the long run matrix has a reduced rank, it has some null eigenvalues, and this should be revealed by the fact that the smallest eigenvalues should have a lot of probability mass accumulated on values close to zero. The computations can be made straightforwardly, simulating values for the long run matrix from its (marginal) posterior distribution, which is a matrix *t*-Student distribution under the non informative prior (16), also considered in the sequel.

Another common procedure is to estimate the rank of Π as the value r that maximizes the (marginal) posterior distribution of the rank. Conditioned on such an estimate, one proceeds to derive the full posterior and eventually estimate the cointegration space, i.e., the linear space spanned by β.

A different approach was proposed by [36], who used the Posterior Information Criterion (PIC), developed in [37], as a criterion to choose the mode of the posterior distribution of the rank of Π. However, as highlighted in [34], one of the advantages of the Bayesian approach is the possibility to incorporate the uncertainty about the parameters in the analysis, represented by the posterior distribution of the rank and, whatever the tool the scientist uses to infer the value of r, it is derived from this posterior distribution.

The authors of [38] nested the reduced rank models in an unrestricted VAR and used Metropolis–Hastings sampling with the Savage–Dickey density ratio—see [39]—to estimate the Bayes Factors of all the models with incomplete rank up to the model with full rank. The Bayes Factor derivation requires the estimation of an error correction factor for the incomplete rank. This factor, however, is not defined for improper priors due to a problem known as *Bartlett paradox*, which arises whenever one compares models of different dimensions. The difficulty is relevant in the present case because, after deriving the rank posterior density, one may consider that models of different dimensions are being compared. The paradox is stated informally as: improper priors should be avoided when one computes Bayes Factors (except for parameters common to both models) as they depend on arbitrary constants (that are integrals).

More recently, [40] developed an efficient procedure to obtain the posterior distribution of the rank using a uniform proper prior over the cointegration space linearly normalized. The author derived solutions for the posterior probabilities for the null rank and for the full rank of Π. The posterior probabilities of each intermediate rank are derived from the posterior samples of the matrices that compose the long run matrix (α and β), properly normalized, under each rank and using the method proposed by [41].

5. Implementing the FBST as a Cointegration Test

This section describes how to implement the FBST to test for cointegration. We will proceed in the same spirit of Section 3, i.e., describing the steps given in Table 1 to implement the test for cointegration.

Let us begin recalling the VECM given by Equation (14):

$$\Delta \mathbf{Y}_t = \mathbf{c} + \Phi_0 \mathbf{D}_t + \Gamma_1 \Delta \mathbf{Y}_{t-1} + \ldots + \Gamma_{p-1} \Delta \mathbf{Y}_{t-p+1} + \Pi \mathbf{Y}_{t-1} + \mathbf{E}_t, \tag{14}$$

$t = 1, \ldots, T + p$, in which $\mathbf{E}_t \overset{i.i.d.}{\sim} N_n(\mathbf{0}, \Sigma)$ with $\mathbf{0}$ a null vector of dimension $n \times 1$ and Ω a symmetric positive definite real matrix. Notice that these assumptions already specify the statistical model (Gaussian) and its implied likelihood. Before giving it explicitly, let us rewrite Equation (14) using matrix notation:

$$\Delta \mathbf{Y} = \mathbf{Z} \cdot \eta + \mathbf{E} \tag{15}$$

where $\Delta \mathbf{Y} = \begin{bmatrix} \Delta \mathbf{Y}'_{p+1} \\ \Delta \mathbf{Y}'_{p+2} \\ \vdots \\ \Delta \mathbf{Y}'_{T+p} \end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix} 1 & \mathbf{D}'_{p+1} & \Delta \mathbf{Y}'_p & \cdots & \Delta \mathbf{Y}'_2 & \mathbf{Y}'_p \\ 1 & \mathbf{D}'_{p+2} & \Delta \mathbf{Y}'_{p+1} & \cdots & \Delta \mathbf{Y}'_3 & \mathbf{Y}'_{p+1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & \mathbf{D}'_{T+p} & \Delta \mathbf{Y}'_{T-1} & \cdots & \Delta \mathbf{Y}'_{T+p-1} & \mathbf{Y}'_{T+p-1} \end{bmatrix}$, $\eta = \begin{bmatrix} \mathbf{c}' \\ \Phi_0 \\ \Gamma_1 \\ \vdots \\ \Gamma_{p-1} \\ \Pi \end{bmatrix}$

and the error vector is given by $\mathbf{E} \sim MN_{T \times n}(0, I_T, \Omega)$, denoting the matrix normal distribution. See Appendix B for more information on this distribution. Now the parameter vector is given by $\Theta = (\eta, \Omega)$.

Notice that $\Delta \mathbf{Y}$ is formed by piling up T transposed vectors $\Delta \mathbf{Y}_t$, thus resulting in a matrix with T lines and n columns (n is the number of time series in vector $\mathbf{Y}_t$), those being also dimensions of matrix $\mathbf{E}$. Matriz $\mathbf{Z}$ is constructed likewise—always piling up the transposed vectors—resulting in a matrix with T lines and $pn + n + 1$ columns. Finally, matrix η has the matrices of coefficients, all piled up properly, resulting in a matrix with $pn + n + 1$ lines and n columns.

Given the assumptions above, $\Delta \mathbf{Y} \sim MN_{T \times n}(\mathbf{Z} \cdot \eta, I_T, \Omega)$, implying that the likelihood is

$$L(\Theta \mid \mathbf{y}) \propto |\Omega|^{-T/2} \exp\left\{ -\frac{1}{2} \cdot \text{tr}[\Omega^{-1}(\Delta \mathbf{Y} - \mathbf{Z} \cdot \eta)'(\Delta \mathbf{Y} - \mathbf{Z} \cdot \eta)] \right\}$$

where $\mathbf{y}$ denotes the set of observed values of vectors $\mathbf{Y}_t$ for $t = 1, \ldots, T + p$. As in Section 3, we will consider an improper prior for Θ, given by

$$h(\Theta) = h(\eta, \Omega) \propto |\Omega|^{-(n+1)/2}, \tag{16}$$

and our reference density, $r(\Theta)$, will be proportional to a constant, leading to a surprise function equivalent to the (full) posterior distribution. These choices correspond to steps 1 and 2 of Table 1. These modeling choices imply the following posterior density:

$$\begin{aligned} g(\Theta \mid \mathbf{y}) &\propto |\Omega|^{-(T+n+1)/2} \exp\left\{ -\frac{1}{2} \cdot \text{tr}[\Omega^{-1}(\Delta \mathbf{Y} - \mathbf{Z} \cdot \eta)'(\Delta \mathbf{Y} - \mathbf{Z} \cdot \eta)] \right\} \\ &= |\Omega|^{-(T+n+1)/2} \exp\left\{ -\frac{1}{2} \cdot \text{tr}\{\Omega^{-1}[\mathbf{S} + (\eta - \widehat{\eta})' \cdot \mathbf{Z}'\mathbf{Z} \cdot (\eta - \widehat{\eta})]\} \right\} \end{aligned} \tag{17}$$

where $\widehat{\eta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\Delta \mathbf{Y}$ and $\mathbf{S} = \Delta \mathbf{Y}'\Delta \mathbf{Y} - \Delta \mathbf{Y}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\Delta \mathbf{Y}$.

To implement Step 3 of Table 1, we need to find the maximum a posteriori of (17) under the constraint $\Theta \subset \Theta_H$, i.e., we need to maximize the posterior in Θ_H. Since we are testing the rank of matrix Π, as discussed in the beginning of Section 4, it is necessary to maximize the posterior assuming the rank of Π is r, $0 \leq r \leq n$. Thanks to the modeling choices made here—Gaussian likelihood and Equation (16) as prior—our posterior is almost identical to a Gaussian likelihood, allowing us to find this maximum using a strategy similar to that proposed by [2], who derived the maximum of the (Gaussian) likelihood function assuming a reduced rank for Π. We will summarize Johansen's algorithm, providing in Appendix C a heuristic argument of why it indeed provides the maximum value of the posterior under the assumed hypotheses.

We begin estimating a VAR$(p-1)$ model for $\Delta\mathbf{Y}_t$ with all the explanatory variables shown in (14) except for $\mathbf{Y}_{t-1}$. Using the matrix notation established above, this corresponds to estimate

$$\Delta\mathbf{Y} = \mathbf{Z}_1 \cdot \eta_1 + \mathbf{U},$$

where $\mathbf{Z}_1 = \begin{bmatrix} 1 & \mathbf{D}'_{p+1} & \Delta\mathbf{Y}'_p & \cdots & \Delta\mathbf{Y}'_2 \\ 1 & \mathbf{D}'_{p+2} & \Delta\mathbf{Y}'_{p+1} & \cdots & \Delta\mathbf{Y}'_3 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \mathbf{D}'_{T+p} & \Delta\mathbf{Y}'_{T-1} & \cdots & \Delta\mathbf{Y}'_{T+p-1} \end{bmatrix}$ and $\eta_1 = \begin{bmatrix} \mathbf{e}' \\ \tau_0 \\ v_1 \\ \vdots \\ v_{p-1} \end{bmatrix}$ showing that $\mathbf{Z}_1$ is obtained

from matrix $\mathbf{Z}$ extracting its last n columns, exactly those corresponding to $\mathbf{Y}_{t-1}$.

We also estimate a second set of auxiliary equations, regressing $\mathbf{Y}_{t-1}$ on a vector of constants and $\mathbf{D}_t, \Delta\mathbf{Y}_{t-1}, \dots, \Delta\mathbf{Y}_{t-p+1}$. By piling up all the (transposed) vectors $\mathbf{Y}'_{t-1}$ for $t = p+1, \dots, T+p$, we have a $(T \times n)$ matrix, denoted by $\mathbf{Y}_{-1}$. As above, these equations can be represented by

$$\mathbf{Y}_{-1} = \mathbf{Z}_1 \cdot \eta_2 + \mathbf{V},$$

where $\mathbf{Y}_{-1} = \begin{bmatrix} \mathbf{Y}'_p \\ \mathbf{Y}'_{p+1} \\ \vdots \\ \mathbf{Y}'_{T+p-1} \end{bmatrix}$ and $\eta_2 = \begin{bmatrix} \mathbf{m}' \\ v_0 \\ \zeta_1 \\ \vdots \\ \zeta_{p-1} \end{bmatrix}$.

Considering the OLS estimates of these sets of equations and their respective estimated residuals, we may write

$$\widehat{\Delta\mathbf{Y}} = \mathbf{Z}_1 \cdot \widehat{\eta}_1 + \widehat{\mathbf{U}} \tag{18}$$

$$\widehat{\mathbf{Y}}_{-1} = \mathbf{Z}_1 \cdot \widehat{\eta}_2 + \widehat{\mathbf{V}} \tag{19}$$

where $\widehat{\eta}_1 = (\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \cdot \Delta\mathbf{Y}$, $\widehat{\eta}_2 = (\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \cdot \mathbf{Y}_{-1}$, $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ are the respective matrices of estimated residuals. Thanks to the Frisch–Waugh–Lovell theorem—see [42] theorem 3.3 or [43] Section 2.4—it is possible to show that the estimated residuals of these auxiliary regressions are related by Π in the following regressions:

$$\widehat{\mathbf{U}} = \Pi\,\widehat{\mathbf{V}} + \widehat{\mathbf{W}}. \tag{20}$$

One can prove that the OLS estimates of Π obtained from (15) and from (20) are numerically identical, as the estimated residuals $\widehat{\mathbf{E}}$ and $\widehat{\mathbf{W}}$.

The second stage of Johansen's algorithm requires the computation of the following sample covariance matrices of the OLS residuals obtained above:

$$\widehat{\Sigma}_{\mathbf{VV}} = \frac{1}{T} \cdot \widehat{\mathbf{V}}'\widehat{\mathbf{V}} \qquad \widehat{\Sigma}_{\mathbf{UU}} = \frac{1}{T} \cdot \widehat{\mathbf{U}}'\widehat{\mathbf{U}}$$

$$\widehat{\Sigma}_{\mathbf{UV}} = \frac{1}{T} \cdot \widehat{\mathbf{U}}'\widehat{\mathbf{V}} \qquad \widehat{\Sigma}_{\mathbf{VU}} = \widehat{\Sigma}'_{\mathbf{UV}}$$

and, from these, we find the n eigenvalues of matrix

$$\widehat{\Sigma}^{-1}_{\mathbf{VV}} \cdot \widehat{\Sigma}_{\mathbf{VU}} \cdot \widehat{\Sigma}^{-1}_{\mathbf{UU}} \cdot \widehat{\Sigma}_{\mathbf{UV}},$$

ordering them decreasingly $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \dots > \widehat{\lambda}_n$. The maximum value attained by the log posterior subject to the constraint that there are r $(0 \le r \le n)$ cointegration relationships is

$$\ell^* = K - \frac{(T+n+1)}{2} \cdot \log|\widehat{\Sigma}_{\mathbf{UU}}| - \frac{T+n+1}{2} \cdot \sum_{i=1}^{r} \log(1-\widehat{\lambda}_i), \tag{21}$$

where K is a constant that depends only on T, n and $\mathbf{y}$ by means of the marginal distribution of the data set, $\mathbf{y}$. Since ℓ^* represents the maximum of the log-posterior, to obtain s^*, one should take $s^* = \exp(\ell^*)$, completing step 3 of Table 1.

As in Section 3, we compute $\overline{ev}$ in step 4 by means of a Monte Carlo algorithm. It is easy to factor the full posterior (17) as a product of a (matrix) normal and an Inverse-Wishart, suggesting a Gibbs sampler to generate random samples from the full posterior. See Appendix B for more on the Inverse-Wishart distribution. Thus, the conditional posteriors for η and Ω are, respectively,

$$g_\eta(\eta \mid \Omega, \mathbf{y}) \propto MN_{n \times k}(\widehat{\eta}, (\mathbf{Z}'\mathbf{Z})^{-1}, \Omega) \tag{22}$$

$$g_\Omega(\Omega \mid \eta, \mathbf{y}) \propto IW(\Omega \mid \mathbf{S} + (\eta - \widehat{\eta})' \cdot \mathbf{Z}'\mathbf{Z} \cdot (\eta - \widehat{\eta}), T) \tag{23}$$

where $\mathbf{S} = \Delta\mathbf{Y}'\Delta\mathbf{Y} - \Delta\mathbf{Y}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\Delta\mathbf{Y}$, IW denotes the Inverse-Wishart, $k = pn + n + 1$ is the number of lines of η, and $\widehat{\eta}$ its OLS estimator, as above. From a Gibbs sampler set with these conditionals, we obtain a random sample from the full posterior to estimate $\overline{ev}$ as the proportion of sampled vectors that generate a value for the posterior greater than s^*. Finally, we obtain $ev = 1 - \overline{ev}$ in the final step (5). The whole implementation for cointegration tests, following the assumptions made in this section, is summarized in Table 4. See Appendix A for more information on the computational resources needed to implement the steps given by Table 4.

Table 4. Pseudocode to implement the FBST to cointegration tests.

General algorithm: compute ev supporting hypothesis $H : \text{rank}(\Pi) = r$ $(0 \leq r \leq n)$ in model (14)
1. Statistical model: Gaussian; prior: $h(\Theta) \propto \|\Omega\|^{-(n+1)/2}$.
2. Reference density: $r(\Theta) \propto 1$; relative surprise function: $g(\Theta \mid \mathbf{y})$.
3. Find s^*: Johansen's algorithm; obtain ℓ^* from Equation (21) with $s^* = \exp(\ell^*)$.
4. Gibbs sampler (from Equations (22) and (23)) to obtain N random samples of parameter vectors from (17). Evaluate the posterior at the sampled vectors and estimate $\overline{ev}$ as the proportion of N for which the evaluated values are larger than s^*.
5. Find $ev = 1 - \overline{ev}$.

Before presenting the results of the procedure applied to real data sets, it is important to remark one feature of the FBST applied to cointegration tests. The estimated eigenvalues of matrix Π, $\widehat{\lambda}_i$, correspond to the squared canonical correlations between $\Delta\mathbf{Y}_t$ and $\mathbf{Y}_{-1}$ corrected for the variable in $\mathbf{Z}_1$ and therefore lie between 0 and 1. Therefore, (21) shows that $\ell_0^* \leq \ell_1^* \leq \ldots \ell_n^*$, where ℓ_r^* denotes the maximum of the posterior (14) assuming Π has rank $0 \leq r \leq n$. Therefore, one may say that the hypotheses $\text{rank}(\Pi) = r$ are nested, in the sense that the respective e-values obtained by the FBST for these hypotheses are always non-decreasing $ev(0) \leq ev(1) \leq \ldots \leq ev(n)$.

This nested formulation is also present in the frequentist procedure proposed by [2], based on the likelihood ratio statistics for successive ranks of Π. Thus, the FBST should be used, like the maximum eigenvalue test, in a sequential procedure to test for the number of cointegrating relationships. We will show how this should be done in presenting the applied results in the sequel.

Results

Now we present, by means of four examples, the application of FBST as a cointegration test. In all the examples, we have adopted a Gaussian likelihood and the improper prior (16). The Gibbs sampler was implemented as described above, providing 51,000 random vectors from the posterior (17). The first 1000 samples were discarded as a burn-in sample, the remaining 50,000 being used to estimate the integral (2). The tables show the e-value computed from the FBST and the maximum eigenvalue test statistics with their respective *p*-values.

Example 1. *We analyzed four electroencephalography (EEG) signals from a subject that has previously presented epileptic seizures. The original study, [44], had the aim of detecting seizures based on multiple hours of recordings for each individual and the cointegration analysis of the mentioned signals was presented by [45]. In fact, the cointegration hypothesis is tested using the phase processes estimated from the original signals. This is done by passing the signal into the Hilbert transform and then "unwrapping" the resulting transform. Sections 2 and 5 of [45,46] provide more details on the Hilbert transform and unwrapping.*

The labels of the modeled series refer to the electrodes on the scalp. As seen in Figures 1 and 2, the series are called FP1-F7, FP1-F3, FP2-F4, and FP2-F8, where FP refers to the frontal lobes and F refers to a row of electrodes placed behind these. Even numbered electrodes are on the right side and odd numbered electrodes are on the left side. The electrodes for these four signals mirror each other on the left and right sides of the scalp. The recordings of the studied subject, an 11-year-old female, identified a seizure in the interval (measured in seconds) [2956, 2996]. Therefore, like [45], we analyze the period of 41 seconds prior to the seizure—interval [2956, 2996]—and the subsequent 41 seconds—interval [2996, 3036]—the seizure period. In the sequel, we will refer to these as *prior to seizure* and *during seizure*, respectively. Since the sample frequency has 256 measurements per second, there are a total of 10,496 measurements for each of the four signals. Ref. [45] used 40 seconds for each period, obtaining slightly different results.

Figures 1 and 2 display the estimated phases based on the original signals. The model proposed by [45] is a VAR(1), resulting in a VECM given by

$$\Delta \mathbf{Y}_t = \mathbf{c} + \Pi \mathbf{Y}_{t-1} + \mathbf{E}_t. \tag{24}$$

Tables 5 and 6 present the results that essentially lead to the same conclusions obtained by [45], even though they have based their findings on the trace test. See Table 8 of [45].

The comparison between *p*-values and the FBST e-values must be made carefully, the main reason being the fact that *p*-values are not measures supporting the null hypothesis, while e-values provide exactly such a kind of support. That being said, a possible way to compare them is by checking the decision their use recommend regarding the hypothesis being tested, i.e., to reject or not the null hypothesis.
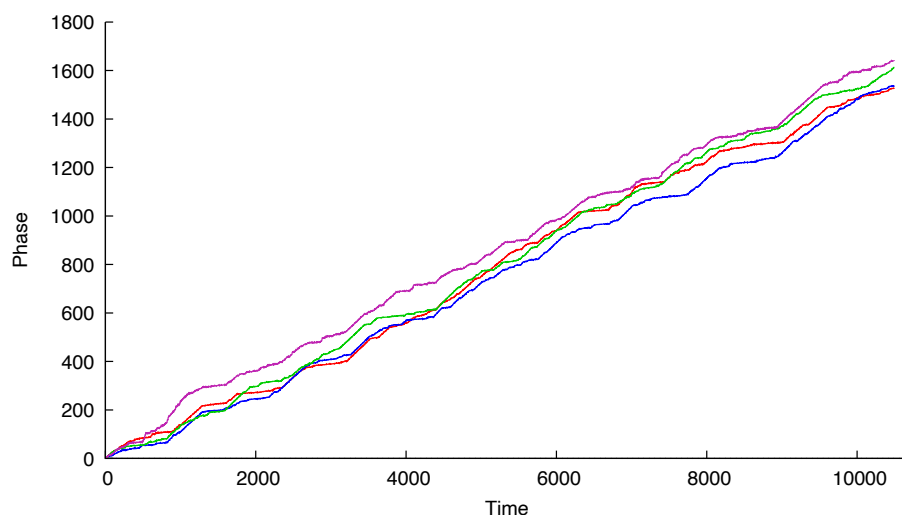

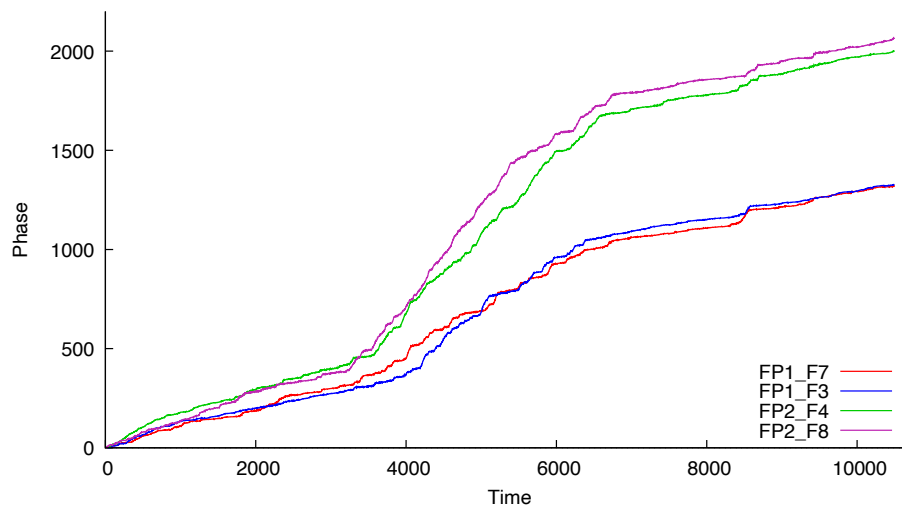**Figure 1.** Estimated phase processes prior to a seizure.

Figure 2. Estimated phase processes during a seizure.

Table 5. FBST and max. eig. test: prior to seizure.

H_0	FBST	Max.	p-Value
$r = 0$	$\simeq 0$	60.966	$\simeq 0$
$r = 1$	0.0691	30.727	0.0010
$r = 2$	0.9990	11.458	0.1337
$r = 3$	$\simeq 1$	0.0812	0.7757

Table 6. FBST and max. eig. test: during seizure.

H_0	FBST	Max.	p-Value
$r = 0$	$\simeq 0$	1120.5	$\simeq 0$
$r = 1$	0.1144	31.563	0.0007
$r = 2$	0.9999	6.5015	0.5574
$r = 3$	$\simeq 1$	1.4383	0.2304

Frequentist tests often adopt a significance level approach: given an observed p-value, the hypothesis is rejected if the p-value is smaller or equal to the mentioned level, usually 0.1, 0.05, or 0.01. Since the cointegration ranks generate nested likelihoods, the hypotheses are tested sequentially, starting with null rank, $r = 0$. For Table 5, adopting a 0.01 significance level, the maximum eigenvalue test would reject $r = 0$ and $r = 1$, and would not reject $r = 2$. The same conclusions follow for Table 6. Thus, the recommended action is to work, for estimation purposes for instance, assuming two cointegration relationships.

The question on which threshold value to adopt for the FBST was already mentioned on Section 1.1, but it is worthwhile to underline it once more. We highly recommend a principled approach deriving the cut-off value from a loss function, which is specific for the problem at hand and the purposes of the analysis. A naive but simpler approach would be to reject the hypothesis if the e-value is smaller than 0.05 or 0.01, emulating the frequentist strategy. Even not recommending this path, since p-values are not supporting measures for the hypothesis being tested while e-values are, the researcher may numerically compare p-values and e-values in a specific scenario. If the researcher derived the p-values from a generalized likelihood ratio test, it is possible to asymptotically compare them. The relationship is: $ev = 1 - F_m[F_{m-h}^{-1}(1 - \mathbf{p})]$, where m is the dimension of the full parameter space, h the dimension of the parameter space under the null hypothesis, F_m the chi-square distribution function with m degrees of freedom and $\mathbf{p}$ the corresponding p-value. See [9,12] for the proof of the asymptotic relationship between e-values and p-values.

Since the maximum eigenvalue test is derived as a likelihood ratio test, this comparison may be done for the results of all the examples presented here, and more appropriately to this example, given its sample size of 10,496 observations. Regarding Tables 5 and 6, one could be in doubt regarding whether to reject or not the hypothesis $r = 1$ since the e-values are larger than 0.01. However, for this model and hypothesis, the e-value corresponding to 0.01 is 0.436. Therefore, in both tables, one could reject the hypothesis and proceed to the next rank that has plenty of evidence in its favor. In conclusion, the practical decisions of both tests (FBST and maximum eigenvalue) would be the same: to not reject $r = 2$.

Example 2 ([47]). *Compare three methods for modeling empirical seasonal temperature forecasts over South America. One of these methods is based on a (possible) long-term cointegration relationship between the temperatures of the quarter March–April–May (MAM) of each year and the temperature of the previous months of November–December–January (NDJ). When there is such a relationship, the authors used the NDJ temperatures (of the previous year) as a predictor for the following MAM season.*

The original data set has monthly temperatures for each coordinate (latitude and longitude) of the covered area. The mentioned series of temperatures (MAM and NDJ) are computed as seasonal averages from this monthly data set by averaging over consecutive three months. Since we have data available from January 1949 to May 2020, the time series of monthly and seasonal average surface temperatures of length 72 for each grid point.

The authors of [47] consider $\mathbf{Y}_t$ as a two-dimensional vector, its first component being the seasonal (average) MAM temperature of year t and the second component the seasonal NDJ temperature of the *previous* year. They consider a VAR(2) without deterministic terms to model the series, resulting in a VECM

$$\Delta \mathbf{Y}_t = \Gamma_1 \Delta \mathbf{Y}_{t-1} + \Pi \mathbf{Y}_{t-1} + \mathbf{E}_t. \tag{25}$$

We have chosen five grid points corresponding to major Brazilian cities to test the cointegration hypothesis of the mentioned seasonal series. The coordinates chosen were the closest ones from: 23.5505° S, 46.6333° W for São Paulo; 22.9068° S, 43.1729° W for Rio de Janeiro; 19.9167° S, 43.9345° W for Belo Horizonte; 15.8267° S, 47.9218° W for Brasília and 12.9777° S, 38.5016° W for Salvador. Figures 3 and 4 show the seasonal temperatures for São Paulo and Brasília, respectively, indicating that the cointegration hypothesis is plausible for both cities.
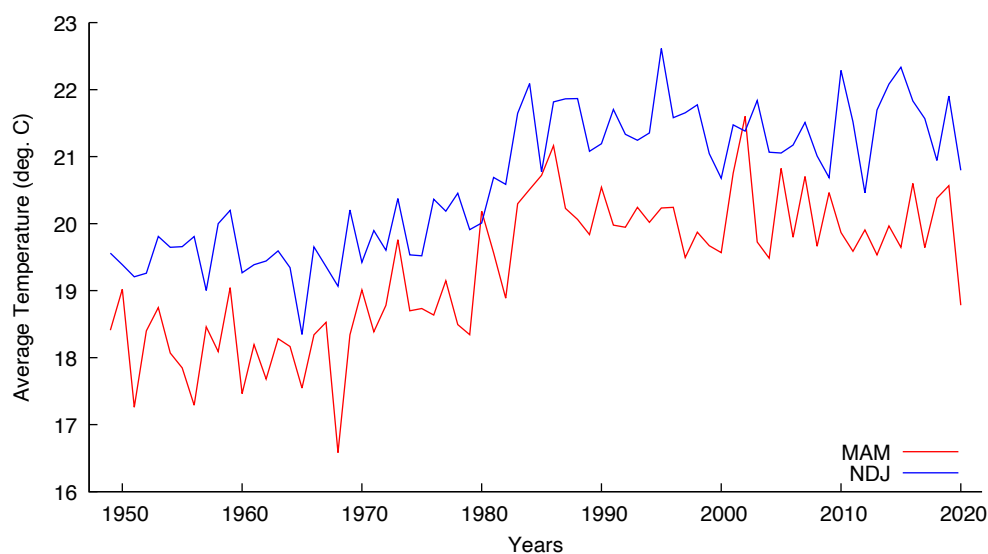

Figure 3. Seasonal (MAM and NDJ) temperatures for São Paulo from 1949 to 2020.
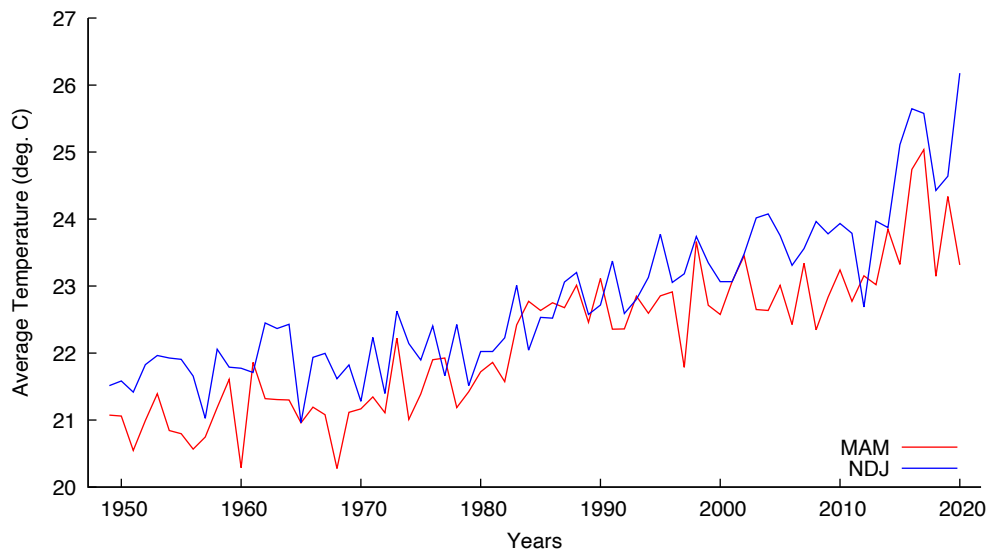
Figure 4. Seasonal (MAM and NDJ) temperatures for Brasília from 1949 to 2020.

Table 7. FBST and maximum eigenvalue test applied to temperature data (MAM and NDJ series) of the mentioned Brazilian cities.

Cities	$H_0 : r = 0$			$H_0 : r = 1$		
	FBST	**Max.**	***p*-Value**	**FBST**	**Max.**	***p*-Value**
São Paulo	0.0012	33.302	$\simeq 0$	$\simeq 1$	0.0893	0.8205
Rio de Janeiro	0.0273	23.294	0.0004	$\simeq 1$	2.43e-5	0.9986
Belo Horizonte	0.0173	24.621	0.0001	$\simeq 1$	0.0963	0.8126
Brasília	0.1129	18.008	0.0045	0.9999	1.3321	0.2892
Salvador	0.0172	24.431	0.0001	$\simeq 1$	0.2450	0.6838

The results are shown in Table 7. Assuming a significance level of 0.01, the maximum eigenvalue test reject the null rank and do not reject $r = 1$ for all the five cities. If we adopt the asymptotic relationship between *p*-values and e-values for the model under analysis, we obtain an e-value of 0.276 corresponding to a 0.01 *p*-value for $r = 0$. Therefore, the FBST would also reject the null rank for all the cities. The hypothesis $r = 1$ is not rejected since all the e-values are close to 1, once more agreeing with the maximum eigenvalue test.

One remark about Brasília seems in order. The city was built to be the federal capital, being officially inaugurated on 21 April 1960. The construction began circa 1957 and before that the site had no human occupation. The process of moving all the administration from Rio de Janeiro, the former capital, was slow and only the 1980 census detected a population over 1 million inhabitants. The present population is almost 3.2 million inhabitants living in the Federal District that includes Brasília and minor surrounding cities. Figure 4 indicates that the seasonal temperatures began to rise exactly after 1980.

Example 3. *we applied the FBST to the Finish data set used in their seminal work [2].*

The authors used the logarithms of the series of *M*1 monetary aggregate, inflation rate, real income, and the primary interest rate set by the Bank of Finland to model the money demand, which, in theory, follows a long-term relationship. The sample has 106 quarterly observations of the mentioned variables, starting at the second trimester of 1958 and finishing in the third trimester of 1984. The chosen model was a VAR(2) with unrestricted constant, meaning that the series in $\mathbf{Y}_t$ have one unit root with drift vector **c** and the cointegrating relations may have a non-zero mean. For more information about how to specify deterministic terms in a VAR see [48], chapter 6. Seasonal dummies for the first three quarters

of the year were also considered in the model chosen by [2]. Writing the model in the error correction form, we have:

$$\Delta \mathbf{Y}_t = \mathbf{c} + \mathbf{\Phi}_{0,1}\mathbf{D}_{1t} + \mathbf{\Phi}_{0,2}\mathbf{D}_{2t} + \mathbf{\Phi}_{0,3}\mathbf{D}_{3t} + \mathbf{\Gamma}_1 \Delta \mathbf{Y}_{t-1} + \mathbf{\Pi} \mathbf{Y}_{t-1} + \mathbf{E}_t. \tag{26}$$

where $\mathbf{\Pi} = \mathbf{\Phi}_1 + \mathbf{\Phi}_2 - I_3$, $\mathbf{\Gamma}_1 = -\mathbf{\Phi}_2$, $\mathbf{c}$ is a vector with constants and $\mathbf{D}_{it}$ denote the seasonal dummies for trimester $i = 1, 2, 3$. The results are displayed in Table 8.

Table 8. FBST and maximum eigenvalue test applied to Finish data of Johansen and Juselius (1990).

H_0	FBST	Max.	p-Value
$r = 0$	0.132	38.489	0.0007
$r = 1$	0.994	26.642	0.0060
$r = 2$	$\simeq 1$	7.8924	0.3983

In [2], the authors concluded that there is, at least, two cointegration vectors, a conclusion that follows if one adopts a 0.01 significance level, for instance. Using the asymptotic relationship between p-values and e-values for Equation (26), we obtain, for $r = 0$, an e-value of 0.998, and, for $r = 1$, an e-value of 0.999, corresponding to a 0.01 p-value. These apparently discrepant values for the e-values are due to the high dimensions of the unrestricted ($m = 58$) and under H_0 ($h = 42$ for $r = 0$ and $h = 43$ for $r = 1$) parameter spaces. Therefore, under this criterion, the FBST also rejects the null rank and $r = 1$ (since $0.132 < 0.998$ and $0.994 < 0.999$, respectively) and does not reject $r = 2$, recommending the same action as the maximum eigenvalue test.

Example 4. *As a final example, we apply the FBST to a US data set discussed in [49]. The observations have annual periodicity and went from 1900 to 1985. We tested for cointegration between real national income, M1 monetary aggregate deflated by the GDP deflator and the commercial papers return rate. The chosen model was a VAR(1) with unrestricted constant. The series were used in natural logarithms and the results follow below:*

Table 9. FBST and maximum eigenvalue test applied to US data of Lucas (2000).

H_0	FBST	Max.	p-Value
$r = 0$	0.042	25.334	0.0101
$r = 1$	0.996	4.2507	0.8271

Table 9 shows that the maximum eigenvalue test rejects $r = 0$ and does not reject $r = 1$ at a 0.05 significance level. Once more adopting the asymptotic relationship between p-values and e-values for the chosen model, we obtain, for $r = 0$, an e-value of 0.247 corresponding to a 0.01 p-value. Thus, under this criterion, the FBST also rejects the null rank and does not reject $r = 1$.

6. Conclusions

In the past few decades, the econometric literature introduced statistical tests to identify unit roots and cointegration relationships in time series. The Bayesian approach applied to these topics advanced considerably after the 1990s, developing interesting alternatives, mostly for unit root testing. The (parametric) frequentist tests mentioned here may not be suitable since these procedures rely on the distribution of the test statistic—usually assuming the hypothesis being tested is true—which depend on a particular a statistical model, usually Gaussian. When the distributions of such statistics cannot be obtained, the procedure is saved by asymptotic results. If the researcher considers different statistical models and the available sample is small, the results of the tests may be quite misleading.

The present work reviewed a simple and powerful Bayesian procedure that can be applied to both purposes: unit root and cointegration testing. We have also shown that the FBST works considerably

well even when one uses improper priors, a choice that may preclude the derivation of Bayes Factors, a standard Bayesian procedure in hypotheses testing.

A long series of articles provided in [7] and the references therein, has showed the versatility and properties of FBST, such as: a. the e-value derivation and computation are straightforward from its general definition; b. it uses absolutely no artificial restrictions like a distinct probability measure on the hypothesis set, induced by some specific parametrization; c. it is in strict compliance with the likelihood principle; d. it can conduct the test with any prior distribution; e. it does not need closed conjectures concerning error distributions, even for small samples; f. it is an exact procedure, since it does rely on asymptotic assumptions; and g. it is invariant with respect to the null hypothesis parametrization and with respect to the parameter space parametrization. See [9], p. 253 for this property.

To proceed with this research agenda, it would be interesting to perform more simulation studies with the FBST applied to unit root testing for a larger group of parametric and semi-parametric models (likelihoods). Another possibility is to include moving average terms in the data generating processes and work with Gaussian and non-Gaussian ARMA models. Notice that, given the points made above, these extensions would not impose major problems to the FBST as they would to the frequentist procedures. Regarding cointegration, the same extensions may be studied in future works, although the adoption of statistical models outside the Gaussian family would require further efforts to numerically implement the FBST. We shall also investigate the effect of the prior choice in the estimates of cointegration relations, especially for small samples.

Appendix A. Computational Resources

The FBST was implemented in all the examples using codes written by the authors in Matlab/ Octave programming language. The results displayed in Tables 3 and 5–9 were obtained using GNU Octave version 4.4.1. The only package required to run the routines was the `statistics` package (version 1.4.1), necessary to simulate vectors of random variables from the distributions mentioned in the text. The codes are briefly described at https://www.ime.usp.br/~jstern/software/, where they can be freely downloaded.

The original data sets used in the examples presented in this work can be obtained from the following sources:

1. Table 3: fourteen U.S. economic time series used by [23]. Available at the R library `urca`, where it is named "npext".
2. Example 1: the original series used in [44,45] are available at https://physionet.org/content/chbmit/1.0.0/. The data for the subject analyzed in Example 1 is from file `chb01_03.edf`, found inside folder `chb01`. To obtain Tables 5 and 6, the data were transformed as described in Example 1.

3. Example 2: the original data set used in [47] is available at https://climexp.knmi.nl/NCEPData/ghcn_cams_05.nc, provided by the Global Historical Climatology Network (GHCN)/Climate Anomaly Monitoring System (CAMS). The data set studied here is the 2 m temperature analysis (0.5 × 0.5) data, a high resolution (0.5 × 0.5 degrees in latitude and longitude) global land surface temperature data set covering the period 1949 to near present, in our case May 2020.
4. Example 3: the original data set with four macroeconomic series used by [2] to estimate the money demand of Finland is available in the R library `urca` with the name "finland".
5. Example 4: the original data used in [49] can be downloaded from https://www.ime.usp.br/~jstern/software/.

Appendix B. Non-Standard Distributions Used in This Article

Appendix B.1. Inverse-Gamma

The probability density function of the Inverse-Gamma distribution is given

$$f_0(x \mid a, b) = \frac{b^a}{\Gamma(a)} \cdot \left(\frac{1}{x}\right)^{a+1} \exp\left(-\frac{b}{x}\right)$$

for $x > 0$ and zero, otherwise. The parameters a and b are both positive real numbers and Γ is the gamma function.

Appendix B.2. Matrix Normal

The probability density function of the random matrix $\mathbf{X}$ with dimensions $p \times q$ that follows the matrix normal distribution $MN_{p \times q}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ has the form:

$$f_1(\mathbf{X} \mid \mathbf{M}, \mathbf{U}, \mathbf{V}) = \frac{\exp\left(-\frac{1}{2}\operatorname{tr}\left[\mathbf{V}^{-1}(\mathbf{X} - \mathbf{M})'\mathbf{U}^{-1}(\mathbf{X} - \mathbf{M})\right]\right)}{(2\pi)^{pq/2}|\mathbf{V}|^{p/2}|\mathbf{U}|^{q/2}}$$

where $\mathbf{M} \in \mathbb{R}^{p \times q}$, $\mathbf{U} \in \mathbb{R}^{p \times p}$ and $\mathbf{V} \in \mathbb{R}^{q \times q}$, being $\mathbf{U}$ and $\mathbf{V}$ symmetric positive semidefinite matrices. The matrix normal distribution can be characterized by the multivariate normal distribution as follows: $\mathbf{X} \sim MN_{p \times q}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ if and only if $\operatorname{vec}(\mathbf{X}) \sim N_{pq}(\operatorname{vec}(\mathbf{M}), \mathbf{V} \otimes \mathbf{U})$, where $\otimes$ denotes the Kronecker product and vec the vectorization of $\mathbf{M}$.

Appendix B.3. Inverse-Wishart

The probability density function of the Inverse-Wishart distribution is

$$f_2(\mathbf{x} \mid \Lambda, \nu) = \frac{|\Lambda|^{\nu/2}}{2^{\nu p/2}\Gamma_p\left(\frac{\nu}{2}\right)} |\mathbf{x}|^{-(\nu+p+1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}(\Lambda \mathbf{x}^{-1})\right]$$

where $\mathbf{x}$ and Λ are $p \times p$ positive-definite matrices, and Γ_p is the multivariate gamma function. Notice that we may also write the same density with $\operatorname{tr}(\mathbf{x}^{-1}\Lambda)$ inside the exponential function, as would be convenient in our implementation of the Gibbs sampler in Section 5.

Appendix C. Heuristic Proof of Johansen's Procedure

The goal of this appendix is to provide a brief heuristic explanation of the procedure, discussed in Section 5 that finds the maximum of posterior (17) subject to the hypothesis that matrix Π has reduced rank r, $0 \le r \le n$. The procedure is based on the algorithm proposed in [2,50] to maximize a Gaussian likelihood under the same assumption (reduced rank of matrix Π). The formal proof of Johansen's algorithm can be found in [51], chapter 20. As mentioned in Section 5, Johansen's algorithm can be applied to the posterior (17) since this distribution is very close to a (multivariate) Gaussian likelihood.

The first step of the algorithm involves "concentrating" the posterior, i.e., to assume Ω and Π are given and maximize the posterior with respect to all the other parameters in Θ. Hence, let γ denote the matrix η except for matrix Π, i.e., $\gamma' = \begin{bmatrix} \mathbf{c} & \Phi_0' & \Gamma_1' & \cdots & \Gamma_{p-1}' \end{bmatrix}$. The concentrated log-posterior, denoted by $\mathcal{M}$, is found by replacing γ with $\widehat{\gamma}(\Pi)$ in (17):

$$\mathcal{M}(\Pi, \Omega \mid \mathbf{y}) = \ln[g(\widehat{\gamma}(\Pi); \Pi, \Omega \mid \mathbf{y})] = C + \frac{(T+n+1)}{2} \ln |\Omega^{-1}| - \left\{ -\frac{1}{2} \cdot \mathrm{tr}[\Omega^{-1}(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}})'(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}})] \right\} \quad \text{(A1)}$$

where C is a constant that depends on T, n and $\mathbf{y}$. The strategy behind concentrating the posterior is that, if we can find the values $\widehat{\Omega}$ and $\widehat{\Pi}$ that maximize $\mathcal{M}$, then these same values, along with $\widehat{\gamma}(\widehat{\Pi})$, will maximize (17) under the constraint $\mathrm{rank}(\Pi) = r$. Carrying the concentration on one step further, we can find the value of Ω that maximizes (A1) assuming Π known, giving

$$\widehat{\Omega}(\Pi) = \frac{1}{T+n+1} \cdot (\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}})'(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}}).$$

To evaluate the concentrated log-posterior at $\widehat{\Omega}(\Pi)$, notice that

$$\mathrm{tr}\left[\widehat{\Omega}(\Pi)^{-1}(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}})'(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}})\right] = \mathrm{tr}[(T+n+1)I_n] = n(T+n+1)$$

and, therefore, denoting by $\mathcal{M}^*$ this new concentrated log-posterior, we have

$$\mathcal{M}^*(\Pi \mid \mathbf{y}) = C + \frac{(T+n+1)n}{2} - \frac{(T+n+1)}{2} \ln \left| \frac{1}{T+n+1}(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}})'(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}}) \right| \quad \text{(A2)}$$

$$= C + \frac{(T+n+1)n}{2} - \frac{(T+n+1)}{2} \ln \left| \frac{T}{T+n+1} \cdot \frac{1}{T}(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}})'(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}}) \right| \quad \text{(A3)}$$

$$= C + \frac{(T+n+1)n}{2} - \frac{(T+n+1)}{2} \ln \left[\left(\frac{T}{T+n+1} \right)^n \cdot \left| \frac{1}{T}(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}})'(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}}) \right| \right] \quad \text{(A4)}$$

$$= K - \frac{(T+n+1)}{2} \cdot \ln \left| \frac{1}{T}(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}})'(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}}) \right| \quad \text{(A5)}$$

where K is a new constant depending only on T, n and $\mathbf{y}$. Equation (A5) represents the maximum value one can achieve for the log-posterior for any given matrix Π. Thus, maximizing the posterior comes down to choosing Π so as to minimize the determinant

$$\left| \frac{1}{T}(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}})'(\widehat{\mathbf{U}} - \Pi\widehat{\mathbf{V}}) \right|$$

subject to the constraint $\mathrm{rank}(\Pi) = r$. The solution of this problem demands the analysis of the sample covariance matrices of the OLS residuals $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ and here we only present the final expression for the maximum value achieved for the log-posterior, denoted ℓ^* in Section 5:

$$\ell^* = K - \frac{(T+n+1)}{2} \cdot \ln |\widehat{\Sigma}_{\mathbf{U}\mathbf{U}}| - \frac{T+n+1}{2} \cdot \sum_{i=1}^{r} \ln(1 - \widehat{\lambda}_i). \quad \text{(A6)}$$

Chapter 20 of [51] provides the formal derivation of (A6).

References

1. Engle, R.F.; Granger, C.W.J. Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica* **1987**, *55*, 251–276. [CrossRef]
2. Johansen, S.; Juselius, K. Maximum likelihood estimation and inference on cointegration—With application to the demand for money. *Oxf. Bull. Econ. Stat.* **1990**, *52*, 169–210. [CrossRef]
3. Diniz, M.A.; Pereira, C.A.B.; Stern, J.M. Unit Roots: Bayesian Significance Test. *Commun. Stats. Theory Methods* **2011**, *40*, 4200–4213. [CrossRef]

4.   Diniz, M.A.; Pereira, C.A.B.; Stern, J.M. Cointegration: Bayesian Significance Test. *Commun. Stats. Theory Methods* **2012**, *41*, 3562–3574. [CrossRef]

5.   Pereira, C.A.B.; Stern, J.M. Evidence and credibility: Full Bayesian Significance Test for precise hypotheses. *Entropy* **1999**, *1*, 69–80.

6.   Pereira, C.A.B.; Stern, J.M.; Wechsler, S. Can a Significance Test Be Genuinely Bayesian. *Bayesian Anal.* **2008**, *1*, 79–100. [CrossRef]

7.   Stern, J.M.; Pereira, C.A.B. The e-value: A Fully Bayesian Significance Measure for Precise Statistical Hypotheses and its Research Program. *São Paulo J. Math. Sci.* **2020**. Available online: https://link.springer.com/article/10.1007%2Fs40863-020-00171-7 (accessed on 20 August 2020).

8.   Good, I.J. *Good Thinking: The Foundations of Probability and Its Applications*; University of Minnesota Press: Minneapolis, MN, USA, 1983.

9.   Stern, J.M. Cognitive Constructivism and the Epistemic Significance of Sharp Statistical Hypotheses in Natural Sciences. *arXiv* **2010**, arXiv:1006.5471. Available online: https://arxiv.org/abs/1006.5471 (accessed on 20 August 2020).

10.   Madruga, M.R.; Esteves, L.G.; Wechsler, S. On the Bayesianity of Pereira-Stern tests. *Test* **2001**, *10*, 291–299. [CrossRef]

11.   Schervish, M. *Theory of Statistics*; Springer: New York, NY, USA, 1995.

12.   Diniz, M.A.; Pereira, C.A.B.; Polpo, A.; Stern, J.M.; Wechsler, S. Relationship between Bayesian and frequentist significance indices. *Int. J. Uncertain. Quantif.* **2012**, *2*, 161–172. [CrossRef]

13.   Borges, W.; Stern, J.M. The rules of logic composition for the Bayesian epistemic E-values. *Log. J. IGPL* **2007**, *15*, 401–420. [CrossRef]

14.   Wald, A. *Statistical Decision Functions*; John Wiley and Sons: New York, NY, USA, 1950.

15.   Tierney, L.; Kadane, J.B. Accurate approximation for posterior moments and marginal densities. *J. Am. Stat. Assoc.* **1986**, *81*, 82–86. [CrossRef]

16.   Dhrymes, P.J. *Mathematics for Econometrics*; Springer: New York, NY, USA, 1978.

17.   Dickey, D.A.; Pantula, S.G. Determining the Ordering of Differencing in Autoregressive Processes. *J. Bus. Econ. Stat.* **1987**, *5*, 455–461.

18.   Sims, C.A. Bayesian skepticism on unit root econometrics. *J. Econ. Dyn. Control* **1988**, *12*, 463–474. [CrossRef]

19.   Sims, C.A.; Uhlig, H. Understanding unit rooters: A helicopter tour. *Econometrica* **1991**, *59*, 1591–1600. [CrossRef]

20.   Phillips, P.C. To criticize the critics: An objective Bayesian analysis of stochastic trends. *J. Appl. Econ.* **1991**, *6*, 333–364. [CrossRef]

21.   Phillips, P.C. Bayesian routes and unit roots: De rebus prioribus semper est disputandum. *J. Appl. Econ.* **1991**, *6*, 435–474. [CrossRef]

22.   Bauwens, L.; Lubrano, M.; Richard, J.-F. *Bayesian Inference in Dynamic Econometric Models*; Oxford University Press: Oxford, UK, 1999.

23.   Schotman, P.C.; van Dijk, H. K. On Bayesian routes to unit roots. *J. Appl. Econ.* **1991**, *49*, 387–401. [CrossRef]

24.   Dickey, D.A.; Fuller, W.A. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **1979**, *74*, 427–431.

25.   Lubrano, M. Testing for unit roots in a Bayesian framework. *J. Econ.* **1995**, *69*, 81–109. [CrossRef]

26.   DeJong, D.; Whiteman, C.H. Reconsidering Trends and random walks in macroeconomic time series. *J. Econ.* **1991**, *28*, 221–254. [CrossRef]

27.   Nelson, C.; Plosser, C. Trends and random walks in macroeconomic time series: Some evidence and implications. *J. Monet. Econ.* **1982**, *10*, 139–162. [CrossRef]

28.   Stern, J.M. Symmetry, Invariance and Ontology in Physics and Statistics. *Symmetry* **2011**, *3*, 611–635. [CrossRef]

29.   MacKinnon, J.G. Approximate asymptotic distribution functions for unit-root and cointegration tests. *J. Bus. Econ. Stat.* **1994**, *12*, 167–176.

30.   Johansen, S. *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*; Oxford University Press: Oxford, UK, 1996.

31.   DeJong, D. Co-integration and trend-stationary in macroeconomic time series. *J. Econ.* **1992**, *52*, 347–370. [CrossRef]

32.   Geweke, J. Bayesian reduced rank regression in econometrics. *J. Econ.* **1996**, *75*, 121–146. [CrossRef]

33. Bauwens, L.; Lubrano, M. *Advances in Econometrics*; JAI Press: Greenwich, CT, USA, 1996.

34. Koop, G.; Strachan, R.; van Dijk, H.K.; Villani, M. *The Palgrave Handbook of Theoretical Econometrics*; Palgrave McMillan: London, UK, 2006.

35. Kleibergen, F.; Paap, R. Priors, posterior odds and Lagrange multiplier statistics in Bayesian analyses of cointegration. *Econ. Inst. Res. Pap.* **1996**. Available online: https://repub.eur.nl/pub/1398/ (accessed on 20 August 2020).

36. Chao, J.; Phillips, P.C. Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. *J. Econ.* **1999**, *91*, 227–271. [CrossRef]

37. Phillips, P.C. Econometric model determination. *Econometrica* **1996**, *59*, 283–306. [CrossRef]

38. Kleibergen, F.; Paap, R. Priors, posterior odds and bayes factors for a Bayesian analysis of cointegration. *J. Econ.* **2002**, *111*, 223–249. [CrossRef]

39. Verdinelli, I.; Wasserman, L. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Am. Stat. Assoc.* **1995**, *90*, 614–618. [CrossRef]

40. Villani, M. Bayesian reference analysis of cointegration. *Econ. Theory* **2005**, *21*, 326–357. [CrossRef]

41. Chib, S.; Greenberg, E. Understanding the Metropolis-Hastings algorithm. *Am. Stat.* **1995**, *49*, 327–335.

42. Greene, W.H. *Econometric Analysis*; Prentice Hall: Bergen County, NJ, USA, 2008.

43. Davidson, R.; MacKinnon, J.G. *Econometric Theory and Methods*; Oxford University Press: Oxford, UK, 2004.

44. Shoeb, A.H. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*; MIT Press: Cambridge, MA, USA, 2009.

45. Østergaard, J.; Rahbeck, A.; Ditlevsen, S. Oscillating systems with cointegrated phase processes. *J. Math. Biol.* **2017**, *75*, 845–883. [CrossRef] [PubMed]

46. Freeman, W.J. Hilbert transform for brain waves. *Scholarpedia* **2007**, *2*, 1338. [CrossRef]

47. Turasie, A.A.; Coelho, C.A.S. Cointegration modeling for empirical South American seasonal temperature forecasts. *Int. J. Climatol.* **2016**, *36*, 4523–4533. [CrossRef]

48. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*; Springer: Berlin, Germany, 2005.

49. Lucas, R. Inflation and welfare. *Econometrica* **2000**, *68*, 247–274. [CrossRef]

50. Johansen, S. Statistical analysis of cointegration vectors. *J. Econ. Dyn. Control* **1988**, *12*, 231–254. [CrossRef]

51. Hamilton, J.D. *Time Series Analysis*; Princeton University Press: Princeton, NJ, USA, 1994.