econometrics

MDPI

Article

Formula I(1) and I(2): Race Tracks for Likelihood Maximization Algorithms of I(1) and I(2) Cointegrated VAR Models

Jurgen A. Doornik [1], Rocco Mosconi [2] and Paolo Paruolo [3,*]

[1] Department of Economics and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford OX1 3UQ, UK; jurgen.doornik@nuffield.ox.ac.uk
[2] Politecnico di Milano, 20133 Milano, Italy; rocco.mosconi@polimi.it
[3] Joint Research Centre, European Commission, 21027 Ispra (VA), Italy
* Correspondence: paolo.paruolo@ec.europa.eu

Academic Editor: Katarina Juselius
Received: 1 July 2017; Accepted: 15 October 2017; Published: 20 November 2017

Abstract: This paper provides some test cases, called circuits, for the evaluation of Gaussian likelihood maximization algorithms of the cointegrated vector autoregressive model. Both I(1) and I(2) models are considered. The performance of algorithms is compared first in terms of *effectiveness*, defined as the ability to find the overall maximum. The next step is to compare their *efficiency* and *reliability* across experiments. The aim of the paper is to commence a collective learning project by the profession on the actual properties of algorithms for cointegrated vector autoregressive model estimation, in order to improve their quality and, as a consequence, also the reliability of empirical research.

Keywords: maximum likelihood; Monte Carlo; VAR; cointegration; I(1); I(2)

JEL Classification: C32; C51; C63; C87; C99

1. Introduction

Since the late 1980s, cointegrated vector autoregressive models (CVAR) have been extensively used to analyze nonstationary macro-economic data with stochastic trends. Estimation of these models often requires numerical optimization, both for stochastic trends integrated of order 1, I(1), and of order 2, I(2). This paper proposes a set of test cases to analyze the properties of the numerical algorithms for likelihood maximization of CVAR models. This is an attempt to start a collective learning project by the profession about the actual properties of algorithms, in order to improve their quality and, as a consequence, the reliability of empirical research using CVAR models.

The statistical analysis of CVAR models for data with I(1) stochastic trends was developed in Johansen (1988, 1991). The I(1) CVAR model is characterized by a reduced rank restriction of the autoregressive impact matrix. Gaussian maximum likelihood estimation (MLE) in this model can be performed by Reduced Rank Regression (RRR, see Anderson 1951), which requires the solution of a generalized eigenvalue problem.

Simple common restrictions on the cointegrating vectors can be estimated explicitly by modifications of RRR, see Johansen and Juselius (1992). However, MLE under more general restrictions, such as equation-by-equation overidentifying restrictions on the cointegration parameters, cannot be reduced to RRR; here several algorithms can be applied to maximize the likelihood. Johansen and Juselius (1994) and Johansen (1995a) provided an algorithm that alternates RRR over each cointegrating vector in turn, keeping the others fixed. They called this a 'switching algorithm', and since then this label has been used for the alternating variables algorithms in the CVAR literature. Boswijk and Doornik (2004) provides an overview.

Switching algorithms have some advantages over quasi-Newton methods: they don't require derivatives, they are easy to implement and each step uses expressions whose numerical properties and accuracy are well known, such as ordinary least squares (OLS), RRR, or generalized least squares (GLS). The downside is that convergence of switching algorithms can be very slow, see Doornik (2017a), and there is a danger of prematurely deciding upon convergence. Doornik (2017a) also showed that adding a line search to switching algorithms can greatly improve their speed and reliability.

The I(2) CVAR is characterized by two reduced rank restrictions, and Gaussian maximum likelihood cannot be reduced to RRR, except in the specific case where it really reduces to an (unrestricted) I(1) model. Initially, estimation was performed by a two-step method, Johansen (1995b). Subsequently, Johansen (1997) proposed a switching algorithm for MLE. Estimation of the I(2) model with restrictions on the cointegration parameters appears harder than in the I(1) case, and it is still under active development, as can be seen below.

While several algorithms exist that estimate the restricted I(1) and I(2) CVAR models, with some of them readily available in software packages, there has been very little research into the effectiveness of these algorithms. No comparative analysis is available either. This paper aims to improve upon this situation; to this effect, it proposes a set of experimental designs that will allow researchers to benefit from the results of alternative algorithms implemented by peers. This should ultimately lead to more effective algorithms, which, in turn, will provide more confidence in the numerical results of empirical analyses.

This paper defines two classes of exercises, called Formula I(1) and I(2), in a playful allusion to Grand Prix car racing championships. Formula I(1) defines a set of precise rules involving I(1) data generation processes (DGP) and models, while Formula I(2) does the same for I(2) DGPs and models. The proposed experiments control for common sources of variability; this improves comparability and efficiency of results. A simple way to control for Monte Carlo variability is to use the same realization of the innovations in the experiments. This is achieved here by sharing a file of innovations and providing instructions on how to build the time series from them.

Econometricians are invited to implement alternative algorithms with respect to the ones employed here, and to test them by running one or more of the exercises proposed in this paper. A companion website https://sites.google.com/view/race-i1 has been created, where researchers interested in checking the performance of their algorithms are invited to follow instruction on how to upload their results, to be compared with constantly-updated benchmarks. Guidelines are illustrated below. The results for all algorithms are by design comparable; moreover, the participation of an additional algorithm may improve the overall confidence in the comparisons, as explained below.

Results from different implementations of algorithms reflect both the properties of the algorithms *sensu stricto* and of their implementation, where one expects different implementations of the same algorithm to lead to different results. Because of this, econometricians are encouraged to participate in the races also with their own implementation of algorithms already entertained by others. This will increase information on the degree of reproducibility of results and on the relative importance of the implementation versus the algorithm *sensu stricto*.[1]

Recent advances in computational technology have fuelled the Reproducible Research movement; the present paper can be seen as a contribution to this movement.[2] The Reproducible Research movement makes use of replication testbeds and Docker containers for replication of results, see e.g., Boettiger (2015). The present project has chosen to keep requirements for researchers at a minimum and it is not demanding the use of these solutions, at least in the current starting configuration.

---

[1] In the rest of the paper the word *algorithm* is used to represent the combination of the algorithm *sensu stricto* and its implementation; hence two implementations of the same algorithm are referred to as two algorithms.

[2] The reference to the Reproducible Research movement was suggested by one referee; see e.g., the bibliography and links at http://reproducibleresearch.net.

The rest of the paper is organized as follows. Section 2 discusses design and evaluation of algorithms in general terms. Section 3 provides definitions, while Section 4 describes precise measures of algorithmic performance. Section 5 describes the Formula I(1) DGP-model pairs, while Section 6 does so for the Formula I(2) DGP-model pairs. The Formula I(1) races are illustrated in Section 7, the Formula I(2) races are illustrated in Section 8; Section 9 concludes. The Appendix A contains practical implementation instructions.

2. Design and evaluation principles

Each exercise in Formula I(1) and I(2) is built around a DGP-model pair. The chosen DGPs have a simple design, with a few coefficients that govern persistence, dimensionality and adjustment towards equilibrium. Aggregating algorithmic performance across runs of the same single DGP appears reasonable, because one expects the frequency of difficult maximization problems to be a function of the characteristics of the given DGP-model pair.

Two main criteria are used for the evaluation of the output from different algorithms. The first one, called *effectiveness*, regards the ability of an algorithm to find a maximum of the likelihood function. Algorithms are expected either to fail, or to converge to a stationary point. This, with some further inspection, may be established as a local maximum. A comparison of local maxima between methods will provide useful insights.

The second one, conditional on the first, is the *efficiency* of the algorithm to find the maximum, which is closely related to its speed. Effectiveness is considered here to be of paramount importance: it is not much use having an algorithm that runs quickly but fails to find the maximum. Actual speeds can be difficult to compare in heterogenous hardware and software environments: however, measures of efficiency can be informative for an implementation in a fixed environment using different designs.

There are many examples of comparisons of optimization algorithms in the numerical analysis literature. Beiranvand et al. (2017) provides an extensive list, together with a generic discussion of how to benchmark optimization algorithms.[3] In the light of this, only advantages and shortcomings of the present approach with respect to Beiranvand et al. (2017) are discussed here, as well as future extensions that are worth considering.

One important specificity here is the focus on the evaluation of estimation procedures for statistical models. These have numerical optimization at their core, but they are applied to maximize specific likelihoods. In the present setting the exact maximum of the likelihood is not known. Moreover, while the asymptotic properties of the MLE are well understood, these will only be approximate at best in any finite sample.

2.1. Race Design

The race design refers to the DGP-model pair, as well as the rules for the implementation of estimators. Because iterative maximization will be used in all cases, algorithms need starting values and decision rules as to when to terminate.

2.1.1. Starting Values

Formula I(1) and I(2) treat the choice of starting value as part of the algorithm. This is the most significant difference with common practice in optimization benchmarking. The starting values may have an important impact on the performance, and, ideally but unfeasibly, one would like to start at the maximum. Optimization benchmarks prescribe specific starting values to create a level playing field for algorithms. This is not done here because implementations may have statistical reasons for

[3] The idea to create public domain test cases based on which programmers may test the performances of their algorithms according to rigorous rules is also not new. For example, the National Institute of Standards and Technology started a website in the late 1990s of the project StRD - Statistical Reference Datasets, see http://www.itl.nist.gov/div898/strd/.

their starting value routine, e.g., there may be a simple estimator that is consistent or an approximation that is reasonable.

Some implementations use a small set of randomized starting values, then pick the best. This approach is general, so could be used by all algorithms. The advantage of the present approach is that one evaluates estimation as it is presented to the user. The drawback is that it will be harder to determine the source of performance differences.[4]

2.1.2. Convergence

The termination decision rule is also left to the algorithm, so it presents a further source of difference between implementations. One expects this to have a small impact: participants in the races should ensure that convergence is tight enough not to change the computed evaluation statistics. If it is set too loose, the algorithm will score worse on reliability. However, setting convergence too tightly will increase the required number of iterations, sometimes substantially if convergence is linear or rounding errors prevent achieving the desired accuracy.

2.1.3. DGP-Model Pair

The chosen DGPs generate I(1) or I(2) processes, as presented in Sections 5 and 6 below; the associated statistical models are (possibly restricted) I(1) or I(2) models, defined in Section 3. Exercises include both cases with correct specification, i.e. when the DGP is contained in the model, as well as cases with mis-specification, i.e., when the DGP is not contained in the model.

Mis-specification is limited here, in the sense that all models still belong to the appropriate model class: an I(1) DGP is always analyzed with an I(1) (sub-)model, and similarly for an I(2) DGP and (sub-)model. Indeed, the sources of mis-specification present here are a subset of the ones faced by econometricians in real applications. The hope is that results for the mis-specification cases covered here can give some lower bounds on the effects of mis-specification for real applications.

Common econometric wisdom says that algorithms tend to be less successful when the model is mis-specified. The present design provides insights as to what extent this is the case in I(1) and I(2) CVAR models, within the limited degree of mis-specification present in these races.

2.1.4. Construction of Test Cases

Different approaches can be used to create test cases:

1. Estimate models on real data

   This is the most realistic setting, because it reflects the complexities of data sets that are used for empirical analyses. On the other hand, it could be hard to study causes of poor performance as there can be many sources such as unmodelled correlations or heteroscedasticities. Aggregating performance over different real datasets may hide heterogeneity in performance due to the different DGPs that have generated the real data.
2. Generate artificial data from models estimated on real data

   This is a semi-realistic setting where it is known from which structure the data are generated. Coefficient matrices of the DGP will normally be dense, with a non-diagonal error variance matrix.
3. Use a purely artificial DGP

   This usually differs from the previous case in that the DGPs are controlled by only a few coefficients that are deemed important. So it is the least realistic case, but offers the possibility to determine the main causes of performance differences.

---

[4] A future extension would be to include some experiments that start from prespecified starting values, as well as storing the initial log-likelihood in the results file.

Formula I(1) and I(2) adopt the artificial DGP approach with sparse design as a method to construct test data. The drawback is that it can only cover a limited number of DGPs, which may not reflect all empirically-relevant situations. The present set of designs is no exception to this rule; however, it improves on the current state of play where no agreed common design of experiments has been proposed in this area. A future extension will be to include a set of tests based on real-world data sets.

2.1.5. Generation of Test Data

All experiments are run with errors that are fixed in advance to ensure that every participant generates exactly the same artificial data sets. The sample size is an important design characteristic; test data are provided for up to 1000 observations, but only races that use 100 or 1000 are included here.

In terms of comparability of results for different algorithms, the possibility to fix the innovations, and hence the data in each lap, controls for one known source of Monte Carlo variability when estimating difference in behavior; see Abadir and Paruolo (2009), Paruolo (2002) or Hendry (1984, §4.1) on the use of common random numbers.

The choice of common random numbers permits to find (significant) differences in behavior of algorithms with a smaller number of cases, and hence computer time, than when innovations vary across teams. Moreover, it also allows the possibility to investigate the presence of multiple maxima.

2.2. Evaluation

Each estimation, unless it ends in failure, results in a set of coefficient estimates with corresponding log-likelihood. Ideally, all algorithms converge to the same stationary point, which is also the global maximum of the likelihood. This will not always happen: it is not known whether these models have unimodal likelihoods, and there is evidence to the contrary in many experiments considered. Moreover, the global maximum is not known, and this is the target of each estimation. As a consequence, evaluation is largely based on a comparison with the best solution.

2.2.1. Effectiveness

The *overall maximum* is the best function value of all algorithms that have been applied to the same problem. This is the best attempt at finding the global maximum, but remains open to revision. Consensus is informative: the more algorithms agree on the maximum, the more confident one is that the global maximum has been found. Similarly, disagreement may indicate multimodality. This is one of the advantages of pooling the results of different algorithms.

If one algorithm finds a lower log-likelihood than another, this indicates that it either found a different local maximum, converged prematurely, or ended up in a saddle point, or, hopefully not so common, there is a programming error. Differences may be the result of the adopted initial parameter values, or owing to the path that is taken, or to the terminal conditions, or a combination of all the above.

However, algorithms that systematically fail to reach the overall maximum should be considered inferior to the ones that find it. Inability to find the global maximum may have serious implications for inference, leading to over-rejection or under-rejection of the null hypothesis for likelihood ratio (LR) tests, depending on whether the maximization error affects the restricted or the unrestricted model.

2.2.2. Efficiency

Efficiency can be expressed in terms of a measure of the number of operations involved, or a measure of time. Time can be expressed as CPU time or as the total time to complete an estimation. While lapsed time is very useful information for a user of the software, it is difficult to use in the present setting. First, the same algorithm implemented in two different languages (say Ox and Matlab) may have very different timings on identical hardware. Next, this project expects submissions of completed results, where the referee team has no control over the hardware.

Even if the referee team were to rerun the experiments, this would be done on different computers with different (versions of) operating systems. Finally, the level of parallelism and number of cores plays a role: even when the considered algorithms cannot be parallelized, matrix operations inside them may be. When running one thousand replications, one could normally do replications in parallel. This suggests not to use time to measure efficiency.

With time measurements ruled out, one is left with counting some aspects of the algorithm. This could be the number of times the objective function is evaluated, the number of parameter update steps, or some other measure. All these measures have a (loose) connection to clock time. E.g., a quadratically convergent algorithm will require fewer function calls and parameter updates than a linearly convergent algorithm, and usually be much faster as well. However, the actual speed advantage can be undermined if the former requires very costly hessian computations (say).

For the switching algorithms that are most commonly used in CVARs when RRR is not possible, the number of parameter update steps is a better metric to express efficiency. An analysis of all the timings reported in Doornik (2017b) shows that, after allowing for two outlying experiments, the number of updates can largely explain CPU time, while the number of objective function evaluations is insignificant. Both the intercept and the coefficient that maps the update count to CPU time are influenced by CPU type, amount of memory, software environment, etc.

In line with common practice, an iteration is defined as a one parameter update step. This definition also applies to quasi-Newton methods, although, unlike switching algorithms, each iteration then also involves the computation of first derivatives. As a consequence, an iteration could be slower than a switching update step, but in many situations a comparison would still be informative. When comparing efficiency of algorithms, the number of iterations appears to be of more fundamental value than CPU time, and it is certainly useful when comparing the same implementation for different experiments when these have been run on a variety of hardware.

There remains one small caveat: changing compiler can affect the number of iterations. When code generation differences mean that rounding errors accumulate differently, this can impact the convergence decision. This effect to be expected to be small.

Summing up, the remainder of the paper uses number of iterations as a measure of efficiency. An update of the parameter vector is understood to define an iteration, and each team participating to Formula I(1) and I(2) is expected to use the same definition.

3. Definitions and Statistical Models

This section introduces the car racing terminology and defines more precisely the notions of DGP and statistical model.

3.1. Terminology

Analogous to car racing terminology, a *circuit* refers to a specific DGP-model pair, i.e. a DGP coupled with a model specification, characterized by given restrictions on the parameters. Each circuit needs to be completed a certain number of times, i.e. *laps* (replications).

Circuits are grouped in two championships, called '*Formula I(1)*' and '*Formula I(2)*'. The implementation of an algorithm corresponds to a driver with a constructor team, which is called a *team* for simplicity. The definition of an algorithm is taken to include everything that it is required to maximize the likelihood function; in particular it includes the choice of the starting value and of the convergence criterion or termination rule.

In the following there are 96 Formula I(1) circuits and 1456 Formula I(2) circuits. Teams do not have to participate in all circuits. For each circuit, a participating team has to:

(i) reconstruct $N = 1000$ datasets (one for each lap) using the innovations provided and the DGP documented below;
(ii) for each dataset, estimate the specified model(s);
(iii) report the results in a given format, described in the Appendix A.

An econometrician may implement more than one algorithm, so enter multiple teams in the races.

3.2. Definitions

This subsection is devoted to more technical definitions of a DGP, a statistical model and its parametrization. A *DGP* is a completely specified stochastic process that generates the sample data $X_{1:T} := (X_1 : \cdots : X_T)$. Here : is used to indicate horizontal concatenation, with the exception for expressions involving indices, such as $(1 : T)$, which is a shorthand for $(1 : \cdots : T)$. For example X_t i.i.d $N(0,1)$, $t = 1, \ldots, T$ is a DGP.

The present design of experiments considers a finite number of DGPs; these are grouped into two classes, called the *I(1) DGP class* and the *I(2) DGP class*. Each DGP class is indexed by a set of coefficients; for example X_t i.i.d $N(0,1)$, $t = 1, \ldots, T$, with $T \in \{100, 1000\}$ is a DGP class.

A parametric *statistical model* is a collection of stochastic processes indexed by a vector of *parameters* φ, which belongs to a *parameter space* Φ, usually an open subset of $\mathbb{R}^m$, where m is the number of parameters. A model is said to be *correctly specified* if its parameter space contains one value of the parameters that characterizes the DGP which has generated the data, and it is mis-specified otherwise. E.g. X_t i.i.d $N(\mu, \sigma^2)$, $t = 1, \ldots, T$, $-\infty < \mu < \infty$, $0 \leq \sigma < \infty$ is a parametric statistical model, when X_t is a scalar and $\varphi = (\mu : \sigma^2)'$. The parameter space Φ is given here by $\mathbb{R} \times \mathbb{R}_+$. Note that a parametric statistical model is needed in order to write the likelihood function for the sample data $X_{1:T}$. In the case above, the likelihood is $f(X_{1:T}; \mu, \sigma^2) = \pi^{-T/2}\sigma^{-T} \exp(-\frac{1}{2}\sum_{t=1}^{T}(X_t - \mu)^2/\sigma^2)$.

Consider now one model A for $X_{1:T}$, indexed by the parameter vector φ with parameter space Φ_A. Assume also that model B is the same, except that the parameter vector φ lies in the parameter space Φ_B with $\Phi_A \subset \Phi_B$. The two models differ by the parameter points that are contained in Φ_B but not in Φ_A (i.e. points in the set $\Phi_B \backslash \Phi_A$). If some points in $\Phi_B \backslash \Phi_A$ cannot be obtained as limits of sequences in Φ_A, (i.e., Φ_B does not coincide with the closure of Φ_A) then model A is said to be a *submodel* of model B. For example, model A can be X_t i.i.d $N(\mu, 1)$, $t = 1, \ldots, T$, $\Phi_A := \{\mu : 0 \leq \mu < \infty\}$ while model B can be $\Phi_B := \{\mu : -\infty < \mu < \infty\}$. Here $\Phi_B \backslash \Phi_A = \{\mu : -\infty < \mu < 0\}$ whose points cannot be obtained as limits of sequences in Φ_A. Hence model A is a submodel of model B.

When all the parameter values in $\Phi_B \backslash \Phi_A$ can be obtained as limits of sequences in Φ_A, then model A and B are essentially the same, and no distinction between them is made here. In this case, or in case the mappings between parametrizations are bijective, it is said that A and B provide *equivalent parametrizations* of the same model. As an example, let $\Phi_A := \{\mu : 0 \leq \mu < \infty\}$ as above and let $\Phi_B := \{\mu = \exp \eta : -\infty < \eta < \infty\}$; the two models are essentially the same, and their parametrizations are equivalent. This is because, despite $\mu = 0$ being present only in the μ parametrization, $\mu = 0$ can be obtained as a limit of points in $\mu = \exp \eta$, $\eta \in (-\infty, \infty)$, e.g., by choosing $\eta_i = -i$, $i = 1, 2, 3 \ldots$ Hence in this case the η and μ parametrizations are equivalent, as they essentially describe the same model.

In the present design of experiments all models are (restricted versions of) the I(1) and the I(2) models, defined below. The case of equivalent models in the above sense is relevant for different parametrizations of the I(2) statistical model, see Noack Jensen (2014).

3.3. The Cointegrated VAR

Both the I(1) and I(2) statistical models are sub-models of the Gaussian VAR model with k lags

$$X_t = \sum_{i=1}^{k} A_i X_{t-i} + \mu_0 + \mu_1 t + \varepsilon_t, \qquad \varepsilon_t \text{ i.i.d. } N(0, \Omega), \quad t = k+1, ..., T, \tag{1}$$

where X_t, ε_t, μ_0, μ_1 are $p \times 1$, A_i and Ω are $p \times p$, and Ω is symmetric and positive definite. The presample values $X_1, ..., X_k$ are fixed and given. The (possibly restricted) parameters associated with μ_0, μ_1, A_i, $i = 1, \ldots, k$ are called the mean-parameters and are indicated by θ. The ones associated with Ω are called the variance parameters, and they are here always unrestricted, except for the

requirement of Ω to be positive definite. The parameter vector is made of the unrestricted entries in θ and Ω.

The Gaussian loglikelihood (excluding a constant term) is given by:

$$-\frac{T-k}{2}\log\det\Omega-\frac{1}{2}\sum_{t=k+1}^{T}\varepsilon_t'(\theta)\,\Omega^{-1}\varepsilon_t(\theta),$$

where $\varepsilon_t(\theta)$ equal to ε_t in (1) considered as a function of θ. Maximizing the loglikelihood with respect to Ω, one finds $\Omega=\Omega(\theta):=(T-k)^{-1}\sum_{t=k+1}^{T}\varepsilon_t(\theta)\varepsilon_t'(\theta)$, which, when substituted back into the loglikelihood gives $-(T-k)p/2+\ell(\theta)$ where

$$\ell(\theta):=-\frac{T-k}{2}\log\det\Omega(\theta),\qquad \Omega(\theta):=\frac{1}{T-k}\sum_{t=k+1}^{T}\varepsilon_t(\theta)\,\varepsilon_t'(\theta). \tag{2}$$

The loglikelihood is here defined as $\ell(\theta)$, calculated as in (2).

The I(1) and I(2) models are submodels of (1).

I(1) statistical models The unrestricted I(1) statistical model under consideration is given by:

$$\Delta X_t=\alpha\beta'\begin{pmatrix}X_{t-1}\\ t\end{pmatrix}+\sum_{i=1}^{k-1}\Gamma_i\Delta X_{t-i}+\mu_0+\varepsilon_t. \tag{3}$$

Here α and $\beta=(\beta^{\star\prime}:\beta_D')'$ are respectively $p\times r$ and $(p+1)\times r$ parameter matrices, $r<p$, with β_D a $1\times r$ vector. The long-run autoregressive matrix $\Pi=-I+\sum_{i=1}^{k}A_i$ is here restricted to satisfy $\operatorname{rank}(\Pi)\le r$, because it is expressed as a product $\Pi=\alpha\beta^{\star\prime}$, where α and $\beta^\star$ have r columns. The coefficient μ_1 is restricted as $\mu_1=\alpha\beta_D'$. The Γ_i matrices are unconstrained. Some Formula I(1) races have restrictions on the columns of α and β.

The I(1) model is indicated as $\mathcal{M}(r)$ in what follows. The likelihood of the I(1) model $\mathcal{M}(r)$ has to be maximized with respect to the parameters α, β, Γ_1, ..., Γ_{k-1}, μ_0 and Ω.

I(2) statistical models The unrestricted I(2) statistical model under consideration is given by:

$$\Delta^2X_t=\alpha\beta'\begin{pmatrix}X_{t-1}\\ t-1\end{pmatrix}+(\Gamma:\mu_0)\begin{pmatrix}\Delta X_{t-i}\\ 1\end{pmatrix}+\sum_{i=1}^{k-2}\Phi_i\Delta^2X_{t-i}+\varepsilon_t, \tag{4}$$

$$\text{with}\qquad \alpha_\perp'(\Gamma:\mu_0)\beta_\perp=\varphi\eta'. \tag{5}$$

Here $\alpha_\perp$ indicates a basis of the orthogonal complement of the space spanned by the columns of α; similarly for $\beta_\perp$ with respect to β. The I(2) model is a submodel of I(1) model; in fact in (4), as in (3), α and $\beta=(\beta^{\star\prime}:\beta_D')'$ are $p\times r$ and $(p+1)\times r$ parameter matrices, $r<p$, with β_D a $1\times r$ vector, and μ_1 is restricted as $\mu_1=\alpha\beta_D'$. In (5), φ is $(p-r)\times s$ and $\eta=(\eta^{\star\prime}:\eta_D')'$ is $(p-r+1)\times s$, $s<p-r$ with η_D a $1\times s$ vector.[5] The Φ_i parameter matrices are unrestricted.

The I(2) model in (4) and (5) is indicated as $\mathcal{M}(r,s)$ in the following. In the I(2) model there are two rank restrictions, namely $\operatorname{rank}(\alpha\beta')\le r$ and $\operatorname{rank}(\alpha_\perp'(\Gamma:\mu_0)\beta_\perp)\le s$. Several different parametrizations exist of the I(2) model $\mathcal{M}(r,s)$, see Johansen (1997), Paruolo and Rahbek (1999),

---

[5] One can observe that $\beta_\perp$ can be chosen as

$$\beta_\perp=\begin{pmatrix}\beta_\perp^\star & \bar\beta^\star\beta_D'\\ 0 & -1\end{pmatrix}$$

so that $\alpha_\perp'(\Gamma:\mu_0)\beta_\perp$ can be written as $\alpha_\perp'(\Gamma:\mu_0)\beta_\perp=(\alpha_\perp'\Gamma\beta_\perp^\star:\alpha_\perp'\Gamma\bar\beta^\star\beta_D'-\alpha_\perp'\mu_0)$. Using the partition $\eta'=(\eta^{\star\prime}:\eta_D')$, Equation (5) can be written as

$$\alpha_\perp'\Gamma\beta_\perp^\star=\varphi\eta^{\star\prime}\qquad\text{and}\qquad\alpha_\perp'\mu_0=\alpha_\perp'\Gamma\bar\beta^\star\beta_D'+\varphi\eta_D'$$

which is the form of the restrictions (5) used in Rahbek et al. (1999) Equation (2.4) (2.5).

Rahbek et al. (1999), Boswijk (2000), Doornik (2017b), Mosconi and Paruolo (2016, 2017), Boswijk and Paruolo (2017). They all satisfy $\text{rank}(\alpha\beta') \leq r$ and $\text{rank}(\alpha'_\perp(\Gamma : \mu_0)\beta_\perp) \leq s$.[6] Teams can choose their preferred parametrization, but, whichever is adopted, the estimated parameters must be reported in terms of α, β, $(\Gamma : \mu_0)$, Φ_1, ..., Φ_{k-2} and Ω.

Some races have restrictions on columns of α, β or τ, where τ is defined as a $(p+1) \times (r+s)$ matrix that spans the column space of $(\beta : \bar{\beta}_\perp \eta)$, where $\bar{a} := a(a'a)^{-1}$.

4. Performance Evaluation

This section defines a number of indicators later employed to measure the performance of algorithms. As introduced above, θ indicates the parameter vector of mean parameters and Ω the variance covariance matrix of the innovations.

4.1. Elementary Information to Be Reported by Each Team

Each lap is indexed by $i = 1, ..., N$, each circuit by $c = 1, \ldots, C$, and each team (i.e., algorithm) by a. The set of teams participating in the race on circuit c is indicated as $\mathcal{A}_c$; this set contains n_c algorithms. The subscript c indicates that $\mathcal{A}_c$ and n_c depend on c, because a team might not participate in all circuits. The following subsections describe the results that each team a has to report, as well as the calculations that the referee team of the race will make on the basis of it.

For each lap i of circuit c, when team a terminates optimization, it produces the optimized value $\theta_{a,c,i}$ of the parameter vector θ. The team should also set the convergence indicator $S_{a,c,i}$ equal to 1 if the algorithm has satisfied the (self-selected) convergence criterion, and set $S_{a,c,i}$ to 0 if no convergence was achieved. Teams should report the loglikelihood value obtained at the maximum $\ell_{a,c,i} := \ell(\theta_{a,c,i})$ using (2).

In case $S_{a,c,i} = 0$, $\theta_{a,c,i}$ indicates the last value of θ before failure of algorithm a. Algorithm a may not have converged either because $\ell(\theta)$ cannot be evaluated numerically anymore (as e.g., when $\Omega(\theta)$ becomes numerically singular) or because a maximum number of iterations has been reached. In the latter case the final loglikelihood should be reported. In the former case, when the likelihood evaluation failed, the team should report $\ell_{a,c,i} = -\infty$.[7] So, regardless of success or failure in convergence, a loglikelihood is always reported.

For each lap i in circuit c, team a should also report the number of performed iterations $N_{a,c,i}$. This number equals the maximum number of iterations if this is the reason why the algorithm terminated. Choosing smaller or larger maximum numbers of iterations will affect result of each team in an obvious way. Teams are asked to choose their own maximum numbers of iterations.

$N_{a,c,i}$ is assumed here to be inversely proportional to the speed of the algorithm. In practice, the speed of the algorithm depends by the average time spent in each iteration, which is influenced by many factors, such as the hardware and software specifications in the implementation. However, because these additional factors vary among teams, the number $N_{a,c,i}$ is taken to provide an approximate indicator of the slowness of the algorithm.

The choice of starting value of the algorithm a is taken to be an integral part of the definition of the algorithm itself. Starting values cannot be based on the results of other teams. It is recommended that the teams document their algorithm in a way that facilitates replication of their results, including providing the computer code used in the calculations and a description of the choice of initial values.

Reported results from the races should be organised in a file, whose name indicates the circuit, and where each row should contain the following information:

$$(i : \ell^u_{a,c,i} : \ell_{a,c,i} : N_{a,c,i} : S_{a,c,i} : \theta^{R\prime}_{a,c,i}),$$

---

[6] In case of no deterministics, the satisfied inequalities are $\text{rank}(\alpha\beta^{\star\prime}) \leq r$ and $\text{rank}(\alpha'_\perp \Gamma \beta^\star_\perp) \leq s$.

[7] Because there is no clear convention on writing $-\infty$, any value of -10^{308} or lower is interpreted as $-\infty$.

where $\ell^u_{a,c,i}$ is the maximum of the loglikelihood of a reference unrestricted model detailed in the Appendix A, and the reported part of the coefficient vector, $\theta^R_{a,c,i}$, is defined as

$$\theta^{R\prime}_{a,c,i} := \left(\mathrm{vec}(\alpha_{a,c,i})' : \mathrm{vec}(\beta_{a,c,i})'\right)$$

for the Formula I(1) circuits. For the Formula I(2) circuits instead:

$$\theta^{R\prime}_{a,c,i} := \left(\mathrm{vec}(\alpha_{a,c,i})' : \mathrm{vec}(\beta_{a,c,i})' : \mathrm{vec}(\Gamma : \mu_0)'_{a,c,i}\right).$$

The reported part of the $\theta_{a,c,i}$ includes the estimated parameters except for the parameters Φ_i of the short term dynamics and the covariance of the innovations Ω. More details on the reporting conventions are given in the Appendix A.

4.2. Indicators of Teams' Performance

After completion of lap i of circuit c by a set of teams $\mathcal{A}_c$, the referee team will compute the overall maximum $\ell^\star_{c,i}$ and deviations $D_{a,c,i}$ from it as:

$$\ell^\star_{c,i} = \max_{\{a\in\mathcal{A}_c:\, S_{a,c,i}=1\}} \ell_{a,c,i}, \qquad D_{a,c,i} = \ell^\star_{c,i} - \ell_{a,c,i}. \tag{6}$$

If all $a \in \mathcal{A}_c$ report failed convergence $S_{a,c,i} = 0$, then $\ell^\star_{c,i}$ will be set equal to $-\infty$ and $D_{a,c,i}$ will be set equal to 0. Observe that $D_{a,c,i} \geq 0$ by construction; $D_{a,c,i}$ is considered small if less than 10^{-7}, moderately small if between 10^{-7} and 10^{-2}, and large if greater than 10^{-2}.[8]

Next define the indicators

$$SC_{a,c,i} := \mathbb{1}(D_{a,c,i} < 10^{-7})S_{a,c,i}, \qquad WC_{a,c,i} := \mathbb{1}(10^{-7} \leq D_{a,c,i} < 10^{-2})S_{a,c,i}$$
$$DC_{a,c,i} := \mathbb{1}(D_{a,c,i} \geq 10^{-2})S_{a,c,i}, \qquad FC_{a,c,i} := 1 - S_{a,c,i},$$

where $\mathbb{1}(\cdot)$ is the indicator function, SC stands for 'strong' convergence, WC stands for 'weak' convergence, DC stands for 'distant' convergence – i.e. convergence to a distant point from the overall maximum – and FC stands for failed convergence. Note that $SC_{a,c,i} + WC_{a,c,i} + DC_{a,c,i} + FC_{a,c,i} = 1$ by construction. When $\ell^\star_{c,i} = -\infty$, note that $SC_{a,c,i} = WC_{a,c,i} = DC_{a,c,i} = S_{a,c,i} = 0$ and $FC_{a,c,i} = 1$.

A summary across laps of the performance of algorithm a in circuit c is given by the quantities

$$SC_{a,c} := 100 \cdot N^{-1} \sum_{i=1}^{N} SC_{a,c,i}, \qquad WC_{a,c} := 100 \cdot N^{-1} \sum_{i=1}^{N} WC_{a,c,i},$$
$$DC_{a,c} := 100 \cdot N^{-1} \sum_{i=1}^{N} DC_{a,c,i}, \qquad FC_{a,c} := 100 \cdot N^{-1} \sum_{i=1}^{N} FC_{a,c,i}.$$

These indicators deliver information on the % of times each algorithm reached strong convergence, weak convergence, convergence to a point which is not the overall maximum, or did not converge.[9]

The set of pairs $\{(\ell^\star_{c,i}, \ell_{a,c,i}) : DC_{a,c,i} = 1\}_{a\in\mathcal{A}_c, i=1:N}$ contain the detailed information on the effects of convergence to a point that is distant from the overall maximum. They are later plotted, together with the distribution of the relevant test statistics. Focusing on the laps where $DC_{a,c,i} = 1$, it is also interesting to calculate the average distance of $\ell_{a,c,i}$ to $\ell^\star_{c,i}$. This is given by

---

[8] As all reference values, the present ones of 10^{-2} and 10^{-7} are chosen in an ad-hoc way. In the opinion of the proponents they reflect reasonable values for the differences between loglikelihoods, which can be interpreted approximately as relative differences. Hence a difference of 10^{-2} means roughly that the two likelihoods differ by 1%, while a difference of 10^{-7} means roughly that the two likelihoods differ by 0.1 in a million, in relative terms.

[9] One may wish to consider these indicators conditionally on the number of converged cases. To do this, one can replace the division by N in the above formulae for $SC_{a,c}$, $WC_{a,c}$, $DC_{a,c}$ with division by $\sum_{i=1}^{N} S_{a,c,i}$.

$$AD_{a,c} = \sum_{i=1}^{N} D_{a,c,i} \cdot DC_{a,c,i} \left(\sum_{i=1}^{N} DC_{a,c,i} \right)^{-1}.$$

Conditionally on convergence, the average number of iterations is defined as

$$IT_{a,c} := \sum_{i=1}^{N} N_{a,c,i} \cdot S_{a,c,i} \left(\sum_{i=1}^{N} S_{a,c,i} \right)^{-1}.$$

4.3. Summary Analysis of Circuits and Laps

The referee team will compute summary statistics for each circuit. First, in order to identify laps where all algorithms fail, the following DNF indicator is defined:

$$DNF_{c,i} = \prod_{a \in \mathcal{A}_c} (1 - S_{a,c,i}),$$

which equals 1 if all algorithms fail to converge.

In order to harvest information on the number of different maxima reported by the teams, the following indicator is constructed. Let $\ell_{(1),c,i} \geq \ell_{(2),c,i} \geq \cdots \geq \ell_{(m_c),c,i}$ be the ordered log-likelihood values reported by those algorithms $a \in \mathcal{A}_c$ that have reported convergence, i.e., for which $S_{a,c,i} = 1$. This list can be used to define the 'number of reported optima' indicator, NOR, as follows

$$NOR_{c,i} = 1 + \sum_{j=2}^{m_c} \mathbb{1}(\ell_{(j-1),c,i} - \ell_{(j),c,i} > 10^{-2}).$$

Note that for each $j = 2, \ldots, m_c$, the difference $\ell_{(j-1),c,i} - \ell_{(j),c,i} \geq 0$ is the decrement of successive reported log-likelihood values; if this decrement is smaller than a selected numerical threshold, here taken to be 10^{-2}, this means that the two algorithms corresponding to $(j-1)$ and (j) have reported the same log-likelihood in practice. In this case, the counter NOR is not increased. If this difference is greater than the numerical threshold of 10^{-2}, then the two reported log-likelihood are classified as different, and the counter NOR is incremented. Overall, NOR counts the number of maxima found by different algorithms that are separated at least by distance of 10^{-2}. NOR is influenced by the number of participating teams.

Observe that the reported log-likelihood value $\ell_{(j),c,i}$ can correspond to an actual maximum or to any other point judged as stationary by the termination criterion used in each algorithm. No check is made by the referee team to distinguish between these situations; NOR should hence be interpreted as an indicator of *potential presence of multiple maxima*; a proper check of the number of maxima would require a more dedicated analysis.[10]

Especially for a difficult lap i, it is interesting to pool information obtained by different algorithms on convergence to points that are distant from the overall maximum, i.e., when $D_{a,c,i}$ is large. This can be averaged across the set of algorithms $\mathcal{A}_c$ that participate to circuit c in the indicator

$$DD_{c,i} = \sum_{a \in \mathcal{A}_c} D_{a,c,i} DC_{a,c,i} \left(\sum_{a \in \mathcal{A}_c} DC_{a,c,i} \right)^{-1}.$$

The indicators $DD_{c,i}$ and $AD_{a,c}$ are obviously related, and they differ in how they are averaged, either across laps or algorithms.

[10] This further analysis is not performed in this paper, but may be considered in later developments of the Formula I(1) and I(2) project.

The DNF and NOR indicators are also aggregated over all laps in a circuit, giving:

$$DNF_c = N^{-1}\sum_{i=1}^{N} DNF_{c,i}, \quad NOR_c = N^{-1}\sum_{i=1}^{N} NOR_{c,i}.$$

5. Formula I(1) Circuits

This section introduces Formula I(1) circuits, i.e. DGP-model pairs where the DGP produces I(1) variables. The model is the I(1) model $\mathcal{M}(r)$ or a submodel of it. For some circuits the statistical model is correctly specified, whereas for others it is mis-specified, as discussed in Section 3.2.

In the specification of the I(1) and I(2) DGPs, the innovations ε_t are chosen uncorrelated (and independent given normality). This choice is made to represent the simplest possible case; this can be changed in future development of the project.[11]

5.1. I(1) DGPs

The I(1) DGP class for lap i is indexed on the scalars (p, T, ρ_0, ρ_1):

$$\begin{cases} \Delta X_{1,t}^{(i)} = \rho_1 \Delta X_{1,t-1}^{(i)} + \varepsilon_{1,t}^{(i)} \\ X_{2,t}^{(i)} = \rho_0 X_{2,t-1}^{(i)} + \varepsilon_{2,t}^{(i)} \end{cases} \qquad \varepsilon_t^{(i)} = \begin{pmatrix} \varepsilon_{1,t}^{(i)} \\ \varepsilon_{2,t}^{(i)} \end{pmatrix} \sim \text{i.i.d. } N\left(0, I_p\right), \tag{7}$$

for $t = 1:T$, $i = 1:N$, $X_t^{(i)} = (X_{1,t}^{(i)\prime} : X_{2,t}^{(i)\prime})'$, $X_0^{(i)} = X_{-1}^{(i)} = 0_p$, where $\varepsilon_{j,t}^{(i)}$ is of dimension $p/2 \times 1$, $j = 1, 2$. Here I_p is the identity matrix of order p. All possible combinations are considered of the following indices and coefficients:

$$p \in \{6, 12\}; \qquad T \in \{100, 1000\}; \qquad \rho_0 \in \{0, 0.9\}; \qquad \rho_1 \in \{0, 0.9\}.$$

Note that in these DGPs,

- the first $p/2$ variables in $X_t^{(i)}$ are either random walks (when $\rho_1 = 0$), or I(1) AR(2) processes whose first difference is persistent (when $\rho_1 = 0.9$). Therefore, ρ_1 is interpreted as 'a near I(2)-ness' coefficient.
- The last $p/2$ variables in $X_t^{(i)}$ are either white noise (when $\rho_0 = 0$), or persistent stationary AR(1) processes (when $\rho_0 = 0.9$). Therefore, ρ_0 is a 'near I(1)-ness' or 'weak mean reversion' coefficient. For simplicity, in the following it is referred to as the 'weak mean reversion' coefficient.

The DGPs can be written as follows, see (3):

$$\Delta X_t = \begin{pmatrix} 0_r \\ (\rho_0 - 1) I_r \end{pmatrix} \begin{pmatrix} 0_r & I_r \end{pmatrix} X_{t-1} + \begin{pmatrix} \rho_1 I_r & 0_r \\ 0_r & 0_r \end{pmatrix} \Delta X_{t-1} + \epsilon_t,$$

where $\mu_0 = \mu_1 = 0$, $r = p/2$, and 0_r is a square block of zeros of dimension r.

To create the Monte Carlo datasets $X_{1:T}^{(i)}$, each team has to use the DGP (7) with relevant values of (p, T, ρ_0, ρ_1), together with the realizations of the ε's as determined by the race organizers. Further details are in the Appendix A.

5.2. I(1) Statistical Models

Using the generated data $X_{1:T}^{(i)}$ as a realization for $X_{1:T}$, the I(1) model $\mathcal{M}(r)$ in (3) has to be estimated on each lap $i = 1, ..., N$; as noted above, MLE of the unrestricted I(1) models $\mathcal{M}(r)$ in (3) is

---

[11] In theory, the chosen DGPs can represent a wider class of DGPs, exploiting invariance of some statistical models with respect to invertible transformation of the variables; see e.g., Paruolo (2005). In practice, however, algorithms may be sensitive to scaling.

obtained by RRR. The estimation sample starts at $t = k+1$, so uses $T-k$ observations. Two alternative values for lag length k are used: $k \in \{2, 5\}$.

All I(1) circuits use the correct rank $r = p/2$ and are subject to further restrictions on the cointegrating vectors, with or without restrictions on their loadings. To express these restrictions, the following matrix structures are introduced, where an $*$ stands for any value, indicating an unrestricted coefficient:

$$\underset{m\times m}{R_{0,m}} = \begin{pmatrix} * & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & * \end{pmatrix}, \quad \underset{m\times m}{R_{1,m}} = \begin{pmatrix} * & 1 & * & * \\ 1 & \ddots & \ddots & * \\ * & \ddots & \ddots & 1 \\ * & * & 1 & * \end{pmatrix}, \quad \underset{m\times m}{R_{2,m}} = \begin{pmatrix} 1 & * & * & * \\ * & \ddots & \ddots & * \\ * & \ddots & \ddots & * \\ * & * & * & 1 \end{pmatrix}. \tag{8}$$

Remark that $R_{0,m}$ sets all elements except the diagonal to zero; $R_{1,m}$ has two bands of unity along the diagonal; $R_{2,m}$ fixes the diagonal to unity, but is otherwise unrestricted. All these matrices are square.

Finally, $U_{m,n}$ stands for an unrestricted $m \times n$ dimensional matrix:

$$\underset{m\times n}{U_{m,n}} = \begin{pmatrix} * & * & ... & * \\ * & * & ... & * \end{pmatrix}. \tag{9}$$

Restriction I(1)-A Model A has the following overidentifying restrictions on β:

$$\beta' = (R_{0,r} : I_r : U_{r,1}). \tag{10}$$

Specification (10) imposes $r(r-1)$ correctly specified overidentifying restrictions on β.

Restriction I(1)-B Model B has over-identifying restrictions that are mis-specified:

$$\beta' = (R_{1,r} : I_r : U_{r,1}). \tag{11}$$

This imposes $2(r-1)$ overidentifying restrictions on β. These restrictions are mis-specified, in the sense that the DGP is outside the parameters space of the statistical model, see Section 3.

Restriction I(1)-C Model C imposes the following, correctly specified, overidentifying restrictions on α and β:

$$\alpha' = (U_{r,r} : R_{0,r}), \qquad \beta' = (R_{0,r} : R_{2,r} : U_{r,1}). \tag{12}$$

Specification (12) imposes $2r(r-1)$ restrictions on α and β. $r(r-1)$ of them would be enough to obtain just-identification, therefore $r(r-1)$ are over-identifying.

6. Formula I(2) Circuits

This section introduces Formula I(2) circuits, following a similar approach to Formula I(1).

6.1. I(2) DGPs

The I(2) DGP class is indexed by the scalars (p, T, ω, ρ_1); the data $X^{(i)}_{1:T}$ is generated as follows:

$$\begin{cases} \Delta^2 X^{(i)}_{1,t} = \varepsilon^{(i)}_{1,t} \\ \Delta X^{(i)}_{2,t} = \rho_1 \Delta X^{(i)}_{2,t-1} + \varepsilon^{(i)}_{2,t} \\ X^{(i)}_{3,t} = \omega X^{(i)}_{3,t-1} + \Delta X^{(i)}_{1,t-1} + \varepsilon^{(i)}_{3,t} \end{cases} \qquad \varepsilon^{(i)}_t = \begin{pmatrix} \varepsilon^{(i)}_{1,t} \\ \varepsilon^{(i)}_{2,t} \\ \varepsilon^{(i)}_{3,t} \end{pmatrix} \sim \text{i.i.d. } N(0, I_p) \tag{13}$$

where $X_0^{(i)} = X_{-1}^{(i)} = 0_p$, $X_t^{(i)} = (X_{1,t}^{(i)\prime} : X_{2,t}^{(i)\prime} : X_{3,t}^{(i)\prime})'$, $t = 1, ..., T$, $i = 1, ..., N$, and $\varepsilon_{j,t}^{(i)}$ is $p/3 \times 1$. As in the I(1) case, all possible combinations are considered of the following indices and coefficients:

$$p \in \{6, 12\}\,; \qquad T \in \{100, 1000\}\,; \qquad \omega \in \{0, 0.9\}\,; \qquad \rho_1 \in \{0, 0.9\}\,.$$

The DGPs can be written as follows, see (4):

$$\Delta^2 X_t = \begin{pmatrix} 0_{2r,r} \\ I_r \end{pmatrix} \begin{pmatrix} 0_{r,2r} & (\omega - 1) I_r \end{pmatrix} X_{t-1} + \begin{pmatrix} 0_r & 0_r & 0_r \\ 0_r & (\rho_1 - 1) I_r & 0_r \\ I_r & 0_r & -I_r \end{pmatrix} \Delta X_{t-1} + \epsilon_t,$$

with $\mu_0 = \mu_1 = 0$. Note that in these DGPs,

- $X_{1,t}$ is a pure cumulated random walk, and hence I(2);
- $X_{2,t}$ is I(1), and does not cointegrate with any other variable in X_t. Moreover, $X_{2,t}$ is a pure random walk when $\rho_1 = 0$, and it is I(1) – near I(2) when $\rho_1 = 0.9$; therefore, as in the I(1) case, the parameter ρ_1 is interpreted as a 'near I(2)-ness' coefficient.
- $X_{3,t}$ is the block of variables that reacts to the multi-cointegration relations, which are given by $(\omega - 1)X_{3,t} + \Delta X_{1,t} - \Delta X_{3,t}$. These relations can be read off as the last block in the equilibrium correction formulation in the last display. When $\omega = 0$ one has that the levels $X_{3,t}$ and differences $\Delta X_{1,t}$, $\Delta X_{3,t}$ have the same weight (apart from the sign) in the multi-cointegration relations; when $\omega = 0.9$ the weight of the levels $1 - \omega = 0.1$ is smaller than the ones of the first differences. Hence ω can be interpreted as the 'relative weight of first differences in the multi-cointegrating relation'.

One can see that in this case:

$$\alpha_\perp = \beta_\perp^\star = \begin{pmatrix} I_r & 0 \\ 0 & I_r \\ 0 & 0 \end{pmatrix},$$

so for the I(2) rank condition:

$$\alpha_\perp' \Gamma \beta_\perp^\star = \begin{pmatrix} 0 & 0 \\ 0 & (\rho_1 - 1) I_r \end{pmatrix} = \begin{pmatrix} 0_r \\ (\rho_1 - 1) I_r \end{pmatrix} \begin{pmatrix} 0_r & I_r \end{pmatrix} = \varphi \eta^{\star\prime}.$$

To create the Monte Carlo dataset $X_{1:T}^{(i)}$, each team has to use the DGP (13) with relevant values of (p, T, ω, ρ_1), together with the drawings of ε as determined by the race organisers. Details are in the Appendix A.

6.2. I(2) Statistical Models

Using the generated data $X_{1:T}^{(i)}$ as a realization for $X_{1:T}$, the I(2) model $\mathcal{M}(r, s)$ in (4) has to be estimated on each lap $i = 1, ..., N$. The estimation sample starts at $t = k + 1$, so uses $T - k$ observations. Two alternative values for the lag k are used, namely $k \in \{2, 5\}$.

An I(2) analysis usually starts with a procedure to determine the rank indices r, s. This requires estimating the $\mathcal{M}(r, s)$ model under all combinations of (r, s), and computing all LR test statistics. Usually, a table is produced with r along the rows and $s_2 = p - r - s$ along the columns.

In the I(2) model $\mathcal{M}(r, s)$, the MLE does not reduce to RRR or OLS except when:

(i) $r = 0$, corresponding to an I(1) model for ΔX_t, or
(ii) $p - r - s = 0$, corresponding to I(1) models for X_t, or
(iii) $r = p$, corresponding to an unrestricted VAR for X_t.

All restricted I(2) circuits use the correct rank indices $r = s = p/3$. Restrictions are expressed using the matrix structures (8) and (9). In addition to restrictions on β and α in (4), there are circuits

with restrictions on τ, which is a basis of the space spanned by $(\beta : \bar{\beta}_{\perp}\eta)$. Under DGP (13), the correctly specified τ is any matrix of the type

$$\tau = \begin{pmatrix} 0_r & 0_r \\ 0_r & I_r \\ I_r & 0_r \\ 0_{1,r} & 0_{1,r} \end{pmatrix} A$$

with $r = s = p/3$ and A any full rank $(r+s) \times (r+s)$ matrix. Recall that 0_r indicates a square matrix of zeros of dimension r; the $0_{1,r}$ vectors are added in the last row to account from the presence of the trend in I(2) model (4).

Unrestricted I(2) models are estimated for

$$1 \leq r \leq p-1, \qquad 0 \leq s \leq p-r-1.$$

The number of models satisfying these inequalities is $p(p-1)/2$. Obviously, some of these models are correctly specified, some are mis-specified.

Restriction I(2)-A Model A is estimated with $r = s = p/3$ under the following overidentifying restrictions on β

$$\beta' = (R_{0,r} : U_{r,r} : I_r : U_{r,1}). \tag{14}$$

This imposes $r(r-1)$ overidentifying restrictions on β. These restrictions are correctly specified.

Restriction I(2)-B The following overidentifying restrictions on β are mis-specified:

$$\beta' = (R_{1,r} : U_{r,r} : I_r : U_{r,1}). \tag{15}$$

This imposes $r(r-1)$ overidentifying restrictions on β, where $r = s = p/3$.

Restriction I(2)-C Overidentifying restrictions on α and β are used in estimation with $r = s = p/3$:

$$\alpha' = (U_{r,2r} : R_{0,r}), \qquad \beta' = (R_{0,r} : U_{r,r} : R_{2,r} : U_{r,1}). \tag{16}$$

Specification (16) imposes $2r(r-1)$ correctly specified restrictions on α and β; $r(r-1)$ of them would be enough to just reach identification of α and β, and hence $r(r-1)$ restrictions are overidentifying.

Restriction I(2)-D Model D has $r = s = p/3$ and $2r(r-1)$ correctly specified overidentifying restrictions on τ of the type:

$$\tau' = \begin{pmatrix} R_{0,r} & I_r & 0_{r,r} & U_{r,1} \\ R_{0,r} & 0_{r,r} & I_r & U_{r,1} \end{pmatrix}. \tag{17}$$

This imposes $2r(r-1)$ overidentifying restrictions on τ.

Restriction I(2)-E The following $2(r-1)+2(s-1)$ mis-specified overidentifying restrictions on τ are imposed in estimation with $r = s = p/3$:

$$\tau' = \begin{pmatrix} R_{1,r} & I_r & 0_{r,r} & U_{r,1} \\ R_{1,r} & 0_{r,r} & I_r & U_{r,1} \end{pmatrix}. \tag{18}$$

7. Test Drive on Formula I(1) Circuits

To illustrate the type of information one can obtain by participating in the Formula I(1) circuits, this Section illustrates a 'test drive' for four algorithms, i.e., teams. The results of these teams also provide a benchmark for other teams willing to participate at a later stage.

Four teams participated in the first Formula I(1) races:

- Team 1: the switching algorithm proposed in Boswijk and Doornik (2004) as implemented in Mosconi and Paruolo (2016) that alternates maximization between β and α. The algorithm is initialized using the unrestricted estimates obtained by RRR. Normalizations are not maintained during optimization, but applied after convergence. The algorithm was implemented in RATS version 9.10.
- Team 2: CATS3 'alpha-beta switching' algorithm as described in Doornik (2017b, §2.2) using the *LBeta* acceleration procedure. CATS3 is an Ox 7 (Doornik (2013)) class for estimation of I(1) and I(2) models, including bootstrapping.
- Team 3: CATS3 'alpha-beta hybrid' algorithm is an enhanced version of alpha-beta switching:

  1. Using standard starting values, as well as twenty randomized starting values, then
  2. alpha-beta switching, followed by
  3. BFGS iteration for a maximum of 200 iterations, followed by
  4. alpha-beta switching.

  This offers some protection against false convergence, because BFGS is based on first derivatives combined with an approximation to the inverse Hessian.

  More important is the randomized search for better starting values as perturbations of the default starting values. Twenty versions of starting values are created this way, and each is followed for ten iterations. Then half are discarded, and they are merged with (almost) identical ones; this is then run for another ten iterations. This is repeated until a single one is left. The iterations used in this start-up process are included in the iteration count.
- Team 4: PcGive algorithm, see Doornik and Hendry (2013, §12.9). This algorithm allows for nonlinear restrictions on α and β, based on switching between the two after a Gauss-Newton warm-up. This is implemented in Ox, Doornik (2013). The iteration count for Team 4 cannot be extracted.

The Formula I(1) circuits are fully described by four features related to the DGP ($p = \{6, 12\}$, $T = \{100, 1000\}, \rho_0 = \{0, 0.9\}, \rho_1 = \{0, 0.9\}$), and two features related to the statistical model: the lag length $k = \{2, 5\}$ and the type of restrictions A, B or C. There are 16 DGPs and 6 model specifications, making a total of 96 circuits.

In the circuits with $T = 1000$, there is not much difference between $k = 2$ and $k = 5$, so the presentation is limited to only one of these values. Combining a long lag length with a small sample size is more problematic. Onatski and Uhlig (2012) consider that situation. They find that the roots of the characteristic polynomial of the VAR tend to a uniform distribution on the unit circle when $\log(T)/k$ and k^3/T tend to zero.

Before analyzing the Formula I(1) races, the tests for cointegration rank are used as 'qualifying races'; this only requires RRR. The qualifying races for Formula I(1) parallel the ones for Formula I(2), reported later. The overall results for the qualifying races show that:

(i) Even when MLE is performed with RRR, inference on the cointegration rank is not easy (not even at $T = 1000$).
(ii) Large VAR dimension, lag length, near-I(2) ness, weak mean reversion are all complicating factors for the use of asymptotic results.

In more detail, Table 1 records the acceptance frequency at 5% significance level of the trace test, using p-values from the Gamma approximation of (Doornik 1998); the null is that the rank is less or equal to r against the alternative of unrestricted rank up to p, where the true rank equals $p/2$. For $T = 1000$ and $p = 6$, the tests behave as expected. When $p = 12$, they tend to favour lower rank values for slow mean-reversion and higher ranks for near-I(2) behaviour.

The results for $T = 100$ are more problematic. When $p = 12$, a lag length of five is excessive relative to the sample size, and leads to overfitting. This is shown in the selection of a too-large rank with frequency close to 1. A lag length of 2 gives opposite results, where a too low rank tends to be

selected away from the near-I(2) cases, and a too high rank is chosen in the near-I(2) cases. In the remainder only $k = 2$ is considered, as this already illustrates an interesting range of results.

Table 1. Formula I(1). Acceptance frequencies at 5% significance level of LR test for rank r against rank p. – indicates exactly zero; other entries rounded to two decimal digits. Bold entries correspond to the true rank $p/2$.

k	T	p	ρ_0, ρ_1	$r = 0$	1	2	3	4	5
2	100	6	0.0, 0.0	–	–	0.20	**0.94**	1.00	1.00
2	100	6	0.0, 0.9	–	–	0.01	**0.57**	0.91	0.99
2	100	6	0.9, 0.0	0.81	0.98	1.00	**1.00**	1.00	1.00
2	100	6	0.9, 0.9	0.18	0.52	0.82	**0.94**	0.99	1.00
5	100	6	0.0, 0.0	0.08	0.41	0.81	**0.96**	1.00	1.00
5	100	6	0.0, 0.9	–	0.03	0.19	**0.52**	0.84	0.97
5	100	6	0.9, 0.0	0.22	0.68	0.92	**0.98**	1.00	1.00
5	100	6	0.9, 0.9	–	0.04	0.23	**0.54**	0.82	0.96
2	1000	6	0.0, 0.0	–	–	–	**0.94**	0.99	1.00
2	1000	6	0.0, 0.9	–	–	–	**0.92**	0.99	1.00
2	1000	6	0.9, 0.0	–	–	0.04	**0.95**	1.00	1.00
2	1000	6	0.9, 0.9	–	–	0.02	**0.93**	0.99	1.00
5	1000	6	0.0, 0.0	–	–	–	**0.94**	1.00	1.00
5	1000	6	0.0, 0.9	–	–	–	**0.92**	0.99	1.00
5	1000	6	0.9, 0.0	–	–	0.13	**0.95**	1.00	1.00
5	1000	6	0.9, 0.9	–	–	0.08	**0.93**	0.99	1.00

k	T	p	ρ_0, ρ_1	$r = 0$	1	2	3	4	5	6	7	8	9	10	11
2	100	12	0.0, 0.0	–	–	0.00	0.06	0.31	0.74	**0.94**	0.98	1.00	1.00	1.00	1.00
2	100	12	0.0, 0.9	–	–	–	–	–	–	**0.01**	0.06	0.19	0.40	0.65	0.90
2	100	12	0.9, 0.0	0.11	0.48	0.81	0.94	0.98	1.00	**1.00**	1.00	1.00	1.00	1.00	1.00
2	100	12	0.9, 0.9	–	–	–	–	0.00	0.02	**0.05**	0.13	0.27	0.42	0.62	0.87
5	100	12	0.0, 0.0	–	–	–	–	0.01	0.05	**0.21**	0.47	0.74	0.90	0.97	0.99
5	100	12	0.0, 0.9	–	–	–	–	–	–	–	–	0.00	0.01	0.06	0.39
5	100	12	0.9, 0.0	–	–	–	–	–	–	**0.00**	0.01	0.05	0.21	0.50	0.82
5	100	12	0.9, 0.9	–	–	–	–	–	–	–	–	–	–	–	0.06
2	1000	12	0.0, 0.0	–	–	–	–	–	–	**0.94**	0.99	1.00	1.00	1.00	1.00
2	1000	12	0.0, 0.9	–	–	–	–	–	–	**0.77**	0.97	0.99	1.00	1.00	1.00
2	1000	12	0.9, 0.0	–	–	0.00	0.01	0.16	0.71	**0.98**	1.00	1.00	1.00	1.00	1.00
2	1000	12	0.9, 0.9	–	–	–	0.00	0.02	0.36	**0.88**	0.98	1.00	1.00	1.00	1.00
5	1000	12	0.0, 0.0	–	–	–	–	–	–	**0.94**	0.99	1.00	1.00	1.00	1.00
5	1000	12	0.0, 0.9	–	–	–	–	–	–	**0.72**	0.96	0.99	1.00	1.00	1.00
5	1000	12	0.9, 0.0	–	–	0.01	0.08	0.39	0.84	**0.98**	1.00	1.00	1.00	1.00	1.00
5	1000	12	0.9, 0.9	–	–	–	0.00	0.11	0.51	**0.87**	0.98	1.00	1.00	1.00	1.00

Table 2 presents the Formula I(1) results for the four teams. For each team, the table reports the convergence quality (SC, WC, DC, for strong, weak, and distant convergence) as percentage of laps, followed by the percentage of laps that failed (FC), the average error distance (AD) and the average iteration count for converged laps only (IT). Team 4 does not report the iteration count. The last two columns are averages over all teams and laps. NOR is the indicator of average number of optima reported, where unity means that in all laps all teams have reported the same maximum.

Table 2. Formula I(1). Selected circuits for 1000 laps. $k = 2$ in all cases. '-' means exactly zero, the other figures are percentages rounded to two decimals and multiplied by 100.

T	p	ρ_0, ρ_1	Team 1						Team 2						Team 3						Team 4						All
			SC	WC	DC	FC	AD	IT	SC	WC	DC	FC	AD	IT	SC	WC	DC	FC	AD	IT	SC	WC	DC	FC	AD	IT	NOR
			Restriction I(1)-A (correctly specified)																								
100	6	0.0,0.0	100	–	–	–	–	16	100	–	–	–	–	4	100	–	–	–	–	4	100	–	–	–	–	–	1
100	6	0.0,0.9	100	–	–	–	–	19	100	–	–	–	–	4	100	–	–	–	–	5	100	–	–	–	–	–	1
100	6	0.9,0.0	92	0	3	5	1	77	95	0	5	–	2	15	96	0	4	–	1	35	96	0	3	1	2	–	1.08
100	12	0.0,0.0	100	–	–	0	–	40	100	–	0	–	3	8	100	–	0	–	3	13	100	–	0	–	2	–	1.00
100	12	0.0,0.9	97	–	2	1	2	74	97	–	3	–	2	15	99	–	1	–	2	50	97	0	2	1	4	–	1.03
100	12	0.9,0.0	72	–	13	15	3	263	77	2	21	–	3	61	85	2	13	–	2	115	77	3	14	7	3	–	1.36
1000	6	0.0,0.0	100	–	–	–	–	6	100	–	–	–	–	1	100	–	–	–	–	2	100	–	–	–	–	–	1
1000	6	0.0,0.9	100	–	–	–	–	6	100	–	–	–	–	1	100	–	–	–	–	2	100	–	–	–	–	–	1
1000	6	0.9,0.0	100	–	–	–	–	11	100	–	–	–	–	3	100	–	–	–	–	4	100	–	–	–	–	–	1
1000	12	0.0,0.0	100	–	–	–	–	7	100	–	–	–	–	2	100	–	–	–	–	3	100	–	–	–	–	–	1
1000	12	0.0,0.9	100	–	–	–	–	7	100	–	–	–	–	2	100	–	–	–	–	3	100	–	–	–	–	–	1
1000	12	0.9,0.0	100	–	–	0	–	20	100	–	–	–	–	5	100	–	–	–	–	5	100	–	–	–	–	–	1
			Restriction I(1)-B (mis-specified)																								
100	6	0.0,0.0	82	–	4	14	2	227	91	1	8	–	5	32	96	1	3	–	4	145	74	0	4	22	2	–	1.12
100	6	0.0,0.9	69	–	8	23	4	336	72	1	28	–	10	84	95	1	4	–	5	225	70	–	8	23	3	–	1.36
100	6	0.9,0.0	81	–	5	13	1	350	81	1	18	–	2	33	86	1	13	–	2	110	71	2	9	19	3	–	1.24
100	12	0.0,0.0	14	–	7	79	3	1839	25	5	69	–	5	790	56	3	40	–	3	1456	10	7	13	70	4	–	1.93
100	12	0.0,0.9	2	–	4	95	9	2286	18	8	73	1	8	1428	66	3	31	–	4	2683	3	2	4	91	5	–	1.89
100	12	0.9,0.0	23	0	10	67	2	2034	28	4	68	–	4	589	45	3	52	–	2	977	10	13	19	57	2	–	1.89
1000	6	0.0,0.0	70	–	2	28	7	546	90	2	8	–	51	32	99	0	1	–	39	136	75	3	2	21	3	–	1.10
1000	6	0.0,0.9	68	0	4	28	41	703	69	1	30	–	163	68	96	1	3	–	127	150	77	0	2	21	102	–	1.35
1000	6	0.9,0.0	85	–	4	11	3	191	93	1	6	–	6	20	96	1	4	–	2	117	76	1	3	21	2	–	1.08
1000	12	0.0,0.0	12	–	10	78	48	1974	18	4	78	1	78	1663	62	3	35	–	104	2434	11	1	3	85	80	–	2.08
1000	12	0.0,0.9	3	–	22	76	79	2542	11	3	80	6	110	1801	66	1	32	2	94	4709	6	–	1	93	62	–	2.18
1000	12	0.9,0.0	22	0	9	68	3	1563	25	6	69	–	5	600	48	4	48	–	4	1177	15	7	14	64	4	–	1.95
			Restriction I(1)-C (correctly specified)																								
100	6	0.0,0.0	99	–	0	0	2	17	100	–	0	–	2	8	100	–	0	–	2	73	94	0	3	3	6	–	1.03
100	6	0.0,0.9	99	–	1	0	2	29	99	–	1	–	2	9	99	–	1	–	1	100	92	–	4	4	5	–	1.05
100	6	0.9,0.0	51	0	15	34	1	396	78	2	20	–	1	55	87	2	10	–	1	182	60	1	15	25	1	–	1.38
100	12	0.0,0.0	65	0	21	15	4	268	68	2	30	–	4	70	71	2	27	–	4	202	59	5	9	28	4	–	1.22
100	12	0.0,0.9	49	–	23	28	6	515	59	3	38	–	5	146	68	3	29	–	5	379	43	9	13	35	6	–	1.35
100	12	0.9,0.0	9	–	13	78	5	2068	38	7	54	0	3	452	63	5	32	–	2	826	16	7	17	60	3	–	1.80
1000	6	0.0,0.0	100	–	–	–	–	6	100	–	–	–	–	4	100	–	–	–	–	89	100	–	–	0	–	–	1
1000	6	0.0,0.9	100	–	–	–	–	6	100	–	–	–	–	4	100	–	–	–	–	107	100	–	–	–	–	–	1
1000	6	0.9,0.0	100	–	–	–	–	11	100	–	–	–	–	6	100	–	–	–	–	30	98	–	1	2	12	–	1.01
1000	12	0.0,0.0	100	–	–	–	–	7	100	–	–	–	–	4	100	–	–	–	–	114	98	–	–	2	–	–	1
1000	12	0.0,0.9	100	–	–	–	–	7	100	–	–	–	–	5	100	–	–	–	–	131	98	–	0	2	160	–	1.00
1000	12	0.9,0.0	95	–	4	1	3	20	95	0	5	–	2	10	95	0	5	–	2	55	72	–	9	19	6	–	1.12

Turning attention to some specific circuits, consider I(1)-A, which has valid overidentifying restrictions on β. For a large enough sample, $T = 1000$, all teams finish equally and quickly. This suggests that any estimation problem for $T = 100$ is a small sample issue.

Consider now the circuits with $p = 6$, $\rho_0 = 0.9$, $\rho_1 = 0$, i.e., the third and ninth row of results in Table 2, panel 1. Figure 1 plots three densities: the empirical densities (kernel density estimates) of the likelihood ratio test $LR^{\star}_{c,i} := 2(\ell^{u}_{c,i} - \ell^{\star}_{c,i})$ for $T = 100, 1000$, where $\ell^{u}_{c,i}$ is the maximized likelihood under the cointegration rank restriction only, along with the $\chi^2(6)$ reference asymptotic distribution. Notice that when $T = 100$ the empirical distribution is very different from the asymptotic one: using the asymptotic 95th percentile of the asymptotic distribution as critical value would lead to severe over-rejection (more than 70%). Finite sample corrections would therefore be very important. Notice that even when $T = 1000$, although the distribution approaches the asymptotic one, the difference is still substantial (the rejection rate is about 10%).

To gain some understanding of the implications of 'distant convergence', for $T = 100$ all laps where any of the teams obtained a distant maximum were pooled, obtaining pairs $\left\{(\ell^{\star}_{c,i}, \ell_{a,c,i}) : DC_{a,c,i} = 1\right\}_{a \in \mathcal{A}_c, i=1:N}$. Figure 1 plots in blue the cdf of the LR test based on the overall maximum, $LR^{\star}_{c,i}$, as the left endpoint of the horizontal lines; the right endpoint represents the LR test based on the distant maximum, i.e., $LR_{a,c,i} = 2(\ell^{u}_{c,i} - \ell_{a,c,i})$. Considering the $\chi^2(6)$, distant convergence has almost no practical implications, since the inappropriate asymptotic distribution would lead to over-rejection anyway. Conversely, relative to the empirical density, in several cases one would (correctly) accept using $LR^{\star}_{c,i}$, and (wrongly) reject using $LR_{a,c,i}$). Distant convergence has therefore implications for hypothesis testing, at least if one takes finite sample problems into account.

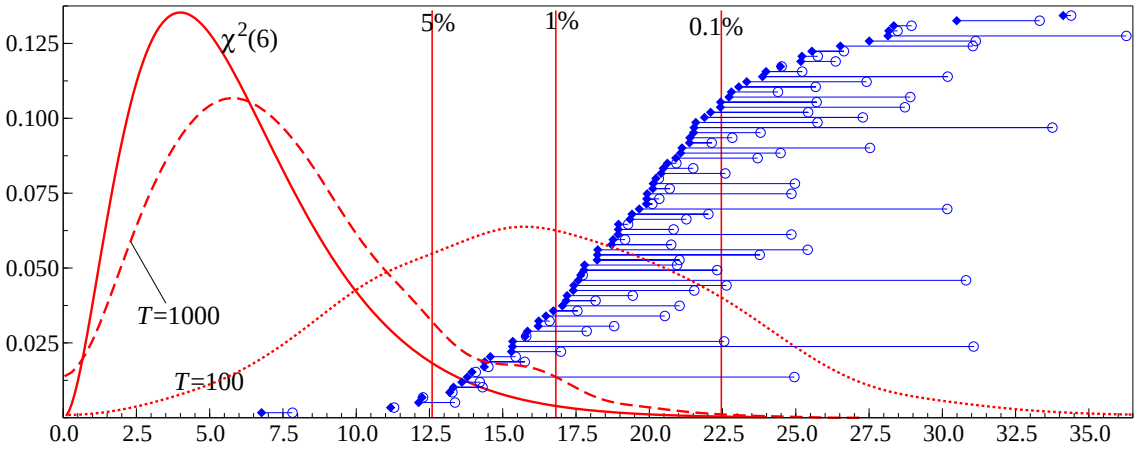

Figure 1. Formula I(1): I(1)-A, $p = 6$, $\rho_0 = 0.9$, $\rho_1 = 0$ circuits. Red: pdfs (on the left scale): kernel-estimate pdfs of $2(\ell^{u}_{c,i} - \ell^{\star}_{c,i})$ for $T = 100$ and $T = 1000$ based on 1000 laps, along with the asymptotic $\chi^2(6)$. Blue: empirical cdf of $2(\ell^{u}_{c,i} - \ell^{\star}_{c,i})$ for $T = 100$, considering only laps and algorithms where distant convergence has been reported. The blue filled diamond denotes the LR calculated using the overall maximum $2(\ell^{u}_{c,i} - \ell^{\star}_{c,i})$, the empty circle the LR calculated using the distant maximum $2(\ell^{u}_{c,i} - \ell_{a,c,i})$.

Consider now the mis-specified restrictions I(1)-B. Table 2 clearly shows that, whichever the DGP, maximizing the likelihood under mis-specified restrictions induces optimization problems. The number of iterations is, for all teams, much higher than under restrictions I(1)-A, and it does not decrease even when $T = 1000$. Failure to converge (FC) becomes a serious problem for teams 1 and 4, whereas teams 2 and 3 do not suffer this problem, but have a much higher percentage of distant convergence (DC). Whether distant convergence is a nuisance or an advantage in this case is however debatable: since the hypothesis is false, rejecting is correct, and therefore distant convergence increases the power of the test (see below).

Figure 2, in analogy with Figure 1, illustrates the challenging 'weak mean reverting' circuits with $p = 6$, $\rho_0 = 0.9$, $\rho_1 = 0$, i.e., the third and ninth row of results in Table 2, panel 3. The asymptotic distribution of the LR test is $\chi^2(4)$, whose 95th percentile is 9.48. Using this as a critical value, $LR^{\star}_{c,i} = 2(\ell^{u}_{c,i} - \ell^{\star}_{c,i})$ would reject about 70% of the times when $T = 100$, and 100% of the times when $T = 1000$. The power seems reasonably good also in small samples, but one needs to keep in mind that, as illustrated when discussing Figure 1, the asymptotic distribution is very inappropriate here.[12]

Figure 2 also illustrates the impact of distant convergence. Using $LR_{a,c,i} = 2(\ell^{u}_{c,i} - \ell_{a,c,i})$ instead of $LR^{\star}_{c,i} = 2(\ell^{u}_{c,i} - \ell^{\star}_{c,i})$ has no practical implication for large samples ($T = 1000$), where the power would be 1 anyway. Conversely, in small samples ($T = 100$) distant convergence has a somewhat beneficial effect on power, increasing the rejection rate. Notice that distant convergence seems to occur more frequently when the null hypothesis is false, like I(1)-B, than when it is true, like I(1)-A; therefore, a tentative optimistic conclusion is that the gain in power due to distant convergence seems to be more relevant than the loss in size.

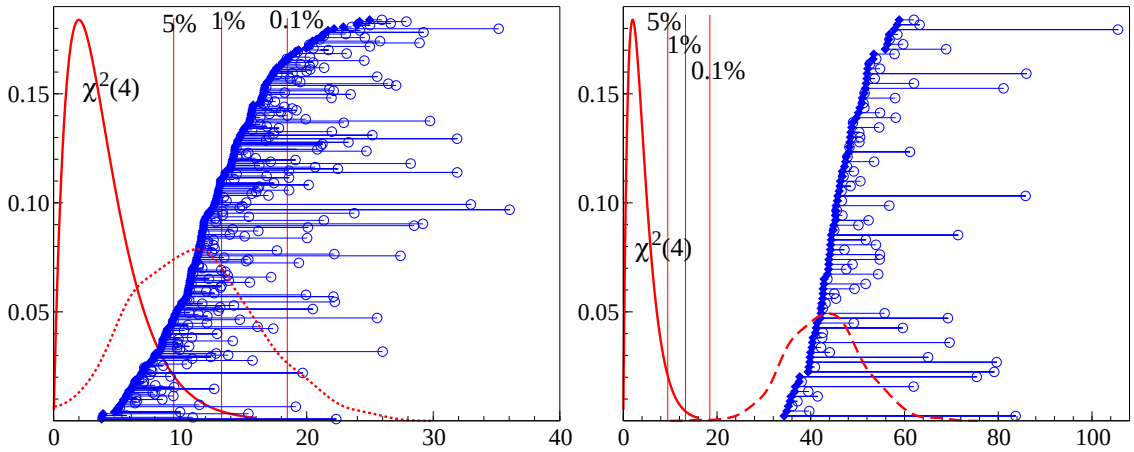

Figure 2. Formula I(1): I(1)-B, $p = 6, \rho_0 = 0.9, \rho_1 = 0$ laps with distant convergence. (**Left**) $T = 100$, (**Right**) $T = 1000$. See caption of Figure 1.

I(1)-C imposes valid over-identifying restrictions on α and β. The first two circuits for I(1)-C are similar to I(1)-A, except that Team 4 has a higher percentage of low and failed convergence. The third circuit shows a more dramatic difference. The effect of the persistent autoregressive effect when $\rho_0 = 0.9$ is to reduce the significance of the α coefficients. As a consequence, some laps yield solutions where some coefficients in α get very large, offset by almost zeros in β. The product Π still looks reasonable, but computation of standard errors of α and β fails (giving huge values), suggesting this may be towards a boundary of the parameter space.

Lap 999 of the third circuit for I(1)-C provides an illustration. Team 1 fails, Teams 2 and 4 have distant convergence. Team 3 has the best results with the following coefficients:

$$\hat{\alpha} = \begin{pmatrix} 0.0369163 & 362.772 & 599.137 \\ -0.00890223 & -17858.0 & -29502.9 \\ 0.0101925 & 11995.8 & 19817.9 \\ -0.0309902 & 0 & 0 \\ 0 & -0.0848045 & 0 \\ 0 & 0 & -0.0545074 \end{pmatrix}, \hat{\beta} = \begin{pmatrix} -4.13543 & 0 & 0 \\ 0 & 1.24557 \cdot 10^{-5} & 0 \\ 0 & 0 & 8.96396 \cdot 10^{-7} \\ 1 & -0.41946 & 0.253898 \\ -6.55434 & 1 & -0.605302 \\ 1.83937 & -1.65209 & 1 \\ 0.664556 & 0.0224937 & -0.013616 \end{pmatrix},$$

[12] Figure 1 shows that, when testing I(1)-A, the asymptotic critical values leads to a 70% rejection rate even if the null hypothesis is true.

where α is numerically close to reduced rank. This model has a loglikelihood that is below the unrestricted I(1) model with rank 2. Because the switching algorithms are really for $\text{rank}(\Pi) \leq r$ rather than $\text{rank}(\Pi) = r$, they occasionally fail or yield unattractive results when α is statistically weakly determined. Team 4 provides more attractive estimates with reasonable standard errors, albeit with a lower loglikelihood.

These characteristics are compatible with several scenarios, including the possibility that in this part of the parameter space the likelihood has a horizontal asymptote. A proper detailed analysis of these and other difficult cases is however beyond the scope of the present paper and it is left for future research.

8. Test Drive on Formula I(2) Circuits

As for Formula I(1), Formula I(2) circuits are illustrated through a test drive for three teams. The following teams participated in the races:

- Team 1: CATS3 'delta switching' algorithm proposed in Doornik (2017b);
- Team 2: CATS3 'triangular switching' algorithm proposed in Doornik (2017b);
- Team 3: CATS3 'tau switching' algorithm proposed in Johansen (1997, §8), implemented as discussed in Doornik (2017b).

As previously illustrated, Formula I(2) is based on 1456 circuits. Although results for all circuits were obtained and stored to serve as benchmark for future comparisons, Formula I(2) circuits and results are too numerous to present in tabular form here; hence only the cases where $p = 6$ and $k = 2$ are shown here.

The first group of circuits are called 'qualifying races', as in Formula I(1). They are designed to:

(i) check the ability of the numerical algorithms to maximize the likelihood of the I(2) model $\mathcal{M}(r, s)$ with no restrictions except for the specification of r and s[13]
(ii) analyze the difficulties of the cointegration ranks tests in spotting the correct r and s in the different DGPs.

Results for task (i) are illustrated in Table 3. For this part of the analysis, only the cases $r = 1, ..., p-1$ and $s = 1, ..., p-r-1$ are considered. This excludes all cases with $r = 0$ and/or $s = p - r$ (i.e., $s_2 = p - r - s = 0$) since in these cases the likelihood of the I(2) model can be maximized exactly by RRR. Preliminary analysis of the results shows that there is relatively little variation for different values of ω and ρ_1. As a consequence, in Table 3 all circuits with the same p, k, T, r, s are analyzed together, irrespective of ω and ρ_1. On the whole, Table 3 shows that the teams perform well. The percentage of 'distant convergence' (DC) is very small, and there are almost no failures. There are a few cases with large 'average distance' (AD), but only when the ranks are smaller than in the DGP.[14] Convergence is quick, usually in about 10 iterations. For $T = 1000$ 'weak convergence' (WC) occurs quite frequently, especially in misspecified (overrestricted) models, and sometimes one observes some large 'average distance' (AD).

---

[13] This aspect of the qualifying races is specific of Formula I(2), since the qualifying races in Formula I(1) can be reduced to RRR.

[14] In the DGP $r = s = s_2 = p/3 = 2$, and the corresponding row is highlighted in boldface in Table 3. See Noack Jensen (2014) for a discussion of the nesting structure of the I(2) models. As illustrated there, all models listed above $r = s = 2$ are misspecified (i.e., overrestricted), whereas all models listed below $r = s = 2$ are correctly specified, since they nest the DGP. Observe that, for example, $r = 2, s = 2$ is nested in $r = 3, s = 0$.

Table 3. Formula I(2). Qualifying races. Performance for I(2) rank-test table, averaged for different values of ρ_1, ω, with a total of 4000 laps. 0: zero to 2 decimals; '-' means exactly zero, the other figures are percentages rounded to two decimals and multiplied by 100.

r,s,s_2	k	T	p	ρ_0,ρ_1	Team 1						Team 2						Team 3							
					SC	WC	DC	FC	AD	IT	SC	WC	DC	FC	AD	IT	SC	WC	DC	FC	AD	IT	DNF	NOR
1,0,5	2	100	6		99	0	1	–	2	13	99	1	1	0	2	17	99	0	1	–	2	13	–	1.02
1,1,4	2	100	6		100	0	0	–	1	12	100	0	0	–	1	12	99	0	1	–	2	13	–	1.01
1,2,3	2	100	6		100	0	0	–	1	12	100	0	0	–	1	12	99	0	0	–	1	14	–	1.00
1,3,2	2	100	6		100	0	–	–	–	9	100	0	–	–	–	8	100	0	–	–	–	11	–	1
1,4,1	2	100	6		100	0	0	–	0	8	100	–	0	–	0	6	100	–	–	0	–	9	–	1.00
2,0,4	2	100	6		100	0	0	–	1	9	100	–	0	0	1	10	99	–	1	–	1	9	–	1.01
2,1,3	2	100	6		100	–	0	–	0	11	100	–	0	–	1	10	100	0	0	–	1	11	–	1.01
2,2,2	**2**	**100**	**6**		**100**	–	**0**	–	0	**6**	**100**	–	**0**	–	**0**	**5**	**100**	**0**	**0**	–	**0**	**8**	–	**1.00**
2,3,1	2	100	6		100	–	0	–	0	5	100	–	0	–	0	4	100	–	–	–	–	6	–	1.00
3,0,3	2	100	6		99	1	0	–	0	12	98	2	0	0	0	15	99	1	0	–	1	12	–	1.00
3,1,2	2	100	6		100	0	0	–	0	10	100	0	0	–	0	11	100	0	0	–	0	10	–	1.00
3,2,1	2	100	6		100	–	0	–	0	8	99	1	0	–	0	16	100	0	0	–	0	8	–	1.00
4,0,2	2	100	6		100	0	0	–	0	9	99	1	0	0	0	18	99	0	0	–	1	10	–	1.00
4,1,1	2	100	6		100	0	0	–	0	9	96	4	0	–	0	22	100	0	0	–	0	8	–	1.00
5,0,1	2	100	6		100	0	0	–	0	6	97	3	0	–	0	18	100	–	0	–	1	7	–	1.00
1,0,5	2	1000	6		78	22	1	–	21	13	46	54	1	–	16	21	78	21	1	–	23	14	–	1.01
1,1,4	2	1000	6		80	20	–	–	–	8	74	26	0	–	0	10	78	22	–	–	–	10	–	1.00
1,2,3	2	1000	6		83	17	–	–	–	8	77	23	–	–	–	9	79	21	–	–	–	11	–	1
1,3,2	2	1000	6		88	12	–	–	–	5	86	15	–	–	–	4	79	21	–	–	–	9	–	1
1,4,1	2	1000	6		87	13	–	–	–	4	87	13	–	–	–	4	77	23	–	–	–	6	–	1
2,0,4	2	1000	6		99	1	0	–	20	3	98	1	0	–	16	3	99	0	1	–	26	3	–	1.01
2,1,3	2	1000	6		96	4	0	–	1	5	92	8	0	–	1	6	95	5	0	–	3	6	–	1.00
2,2,2	**2**	**1000**	**6**		**100**	–	–	–	–	**1**	**100**	–	–	–	–	**1**	**100**	–	–	–	–	**1**	–	**1**
2,3,1	2	1000	6		100	0	–	–	–	1	100	0	–	–	–	1	100	0	–	–	–	1	–	1
3,0,3	2	1000	6		81	19	0	–	0	10	62	38	0	–	0	16	80	19	0	–	0	10	–	1.00
3,1,2	2	1000	6		91	8	1	–	1	11	91	8	1	–	1	13	99	1	0	–	1	7	–	1.01
3,2,1	2	1000	6		97	3	0	–	0	8	67	33	0	–	0	18	100	0	0	–	0	5	–	1.00
4,0,2	2	1000	6		97	3	0	–	0	9	74	26	0	0	0	26	97	2	0	–	1	10	–	1.01
4,1,1	2	1000	6		98	2	0	–	0	8	48	52	0	0	0	33	100	0	0	–	0	6	–	1.01
5,0,1	2	1000	6		100	0	–	–	–	6	48	52	0	0	0	29	99	0	0	–	0	6	–	1.00

On the whole, likelihood maximization is reasonably accurate, for each circuit and lap; one can then proceed to find the maximum of the maximized likelihoods reported by the three teams. On this basis, the likelihood ratio tests for the cointegration ranks r and s were computed on the overall maximum. As done in Table 1 for the I(1) case, Table 4 records the acceptance frequency at 5% significance level of the LR cointegration test, using p-values from the Gamma approximation of (Doornik 1998); the null is that rank $(\alpha\beta') \leq r$ and rank $(\alpha'_\perp(\Gamma : \mu_0)\beta_\perp) \leq s$ against the alternative of unrestricted VAR.

Table 4. Formula I(2). Qualifying race. I(2) rank-test table, $p = 6, k = 2, r = 1$. Acceptance frequencies at 5% significance level of LR test for ranks (r, s) against the unrestricted VAR, with $p = 6$, $k = 2$ and $s_2 = p - r - s$. Bold entries correspond to the true ranks $r = s = s_2 = p/3 = 2$. Cases with $r = 0$ and $r = 1$ have been omitted for readability, since the acceptance rate is always zero or very close to zero.

T	ρ_0, ρ_1	$r = 2$					$r = 3$				$r = 4$			$r = 5$	
		$s_2 = 4$	3	2	1	0	$s_2 = 3$	2	1	0	$s_2 = 2$	1	0	$s_2 = 1$	0
100	0.0, 0.0	0.00	0.01	**0.87**	0.69	0.34	0.88	0.97	0.94	0.74	0.98	0.99	0.92	0.99	0.99
100	0.0, 0.9	0.91	0.87	**0.65**	0.32	0.07	0.98	0.93	0.73	0.34	0.98	0.93	0.69	0.98	0.93
100	0.9, 0.0	0.01	0.25	**0.50**	0.28	0.06	0.52	0.76	0.66	0.33	0.83	0.87	0.64	0.92	0.91
100	0.9, 0.9	0.58	0.46	**0.25**	0.07	0.01	0.74	0.62	0.32	0.09	0.82	0.67	0.34	0.87	0.72
1000	0.0, 0.0	0.00	0.00	**0.94**	0.85	0.48	0.94	0.99	0.98	0.83	0.99	1.00	0.95	1.00	1.00
1000	0.0, 0.9	0.00	0.12	**0.93**	0.77	0.40	0.94	0.99	0.97	0.78	0.99	1.00	0.94	1.00	0.99
1000	0.9, 0.0	0.00	0.00	**0.92**	0.77	0.39	0.90	0.98	0.97	0.77	0.99	0.99	0.93	1.00	0.99
1000	0.9, 0.9	0.00	0.07	**0.88**	0.70	0.33	0.91	0.98	0.95	0.71	0.99	0.99	0.91	1.00	0.98

For this aspect of the analysis, all cases $r = 0, ..., p - 1$ and $s = 1, ..., p - r$ are considered. However, Table 4 does not report $r = 0$, since the acceptance rate is exactly zero for all values of s in that case. Also the case $r = 1$ is not reported, because the acceptance rate is always zero for $T = 1000$ and very close to zero for $T = 100$ (less than 0.02, except for $\omega = \rho_1 = 0.9$, where it is 0.06). The model corresponding to the DGP, i.e., $r = s = s_2 = p/3 = 2$, has been highlighted in boldface.

Note that r is almost never underestimated even when $T = 100$, irrespective of the value of ω. This seems to be a major difference with respect to Formula I(1), where r is frequently underestimated when ρ_0 is 0.9. It is important to remark that the interpretation of ρ_0 in Formula I(1) different from the interpretation of ω in Formula I(2), although they both affect the magnitude of the coefficients in Π. In fact ρ_0 may be interpreted as 'weak mean reversion', whereas ω has no implication for the speed of adjustment, but it is rather related to the relative weight of levels and differences in the polynomial cointegration relations; this might be the reason why for $\omega = 0.9$ there is no to underestimation of r.[15] It is, however, surprising that when $\omega = 0.9$ (so that the weight of the levels is reduced to $1 - \omega = 0.1$) one tends to overestimate r, rejecting $r = 2$ in favour of $r = 3$ or even $r = 4$.

The impact of the 'near I(2)' parameter ρ_1 is linked to the form of the DGP in (13): when $\rho_1 = 0.9$ the variables in $\Delta X_{2,t}$ are stationary but slowly mean reverting, so that $X_{2,t}$ is almost I(2). Not surprisingly then, when $\rho_1 = 0.9$ the tests tend to underestimate s (i.e., overestimate $s_2 = p - r - s$) at least when $T = 100$, so that very frequently $r = 2, s = 0$ is selected. When $T = 1000$ the power vs $s = 0$ goes to 1, but one would still select $r = 2, s = 1$ about 10% of the times.

The results on the Formula I(2) circuits with restrictions on the cointegration parameters (in addition to the restrictions on the ranks) are illustrated in Table 5. The cases I(2)-A, I(2)-B and I(2)-C involve only the matrix Π; more specifically, as in Formula I(1), models I(2)-A involve correctly

[15] Formula I(2) circuits may be extended in the future introducing another coefficient in analogy with ρ_0 of Formula I(1). This would amount at replacing the third equation in (13) with

$$\Delta^2 X_{3,t}^{(i)} = (\rho_0 - 1)((\omega - 1)X_{3,t-1}^{(i)} + \Delta X_{1,t-1}^{(i)} - \Delta X_{3,t-1}^{(i)}) + \varepsilon_{3,t}^{(i)}.$$

specified restrictions on β, models I(2)-B contain misspecified restrictions on β, while models I(2)-C contain correctly specified restriction on α and β. All three algorithms seem to perform quite well in maximizing the likelihood of the I(2) model under restrictions on Π only, with Triangular-hybrid beating the others. In particular, under the correctly specified restrictions A and C the likelihood is easily and quickly maximized (especially when $T = 1000$), with almost no case of distant convergence.

Conversely, the misspecified restrictions in model I(2)-B require more iterations and, for the first two teams, induce distant convergence quite frequently. However, as observed when discussing Formula I(1) results, it is important to keep in mind that distant convergence is indeed a problem when the restriction is correctly specified since it leads to over-rejection, whereas for misspecified restrictions it can be seen as beneficial, since it increases the power of the test.

More generally, the analysis of restrictions A, B, C, seems to suggest that estimation of restricted α and β is easier in the I(2) case with respect to the I(1) case. Note however that the comparison is not completely fair, since most of the difficulties in the I(1) case are found when $\rho_0 = 0.9$ (weak mean reversion), and this coefficient does not appear in the current Formula I(2) design.

Consider finally the restrictions I(2)-D and I(2)-E, reported in the last two panels of Table 5. Remember that I(2)-D is a correctly specified model with restrictions on τ, while I(2)-E is a misspecified model with restrictions on τ. Table 5 shows serious difficulties in maximizing the likelihood under restrictions on τ; in both cases (i.e., whether the hypothesis is true or false), the number of iterations is much higher than under restrictions A, B and C and it does not decrease even when $T = 1000$. Failure to converge (FC) becomes a serious problem for triangular switching (and to some extent delta switching), and there is an high percentage of distant convergence (DC) for all three algorithms; Triangular hybrid performs better, having a smaller average distance (AD). Notice that for model I(2)-D (where the null hypothesis is true) distant convergence is more problematic since it leads to over-rejection.

To analyze this problem, as done in Formula I(1), Figure 3 illustrates the impact of distant convergence. It is apparent from the figure that over-rejection is substantial here. Since the 5% critical value of the asymptotic $\chi^2(4)$ distribution is 9.49, the analysis clearly shows several cases where one would (correctly) accept using the overall maximum, and (wrongly) reject using the distant maximum. The striking difference with respect to Formula I(1) is that here the over-rejection due to distant convergence remains even when $T = 1000$.

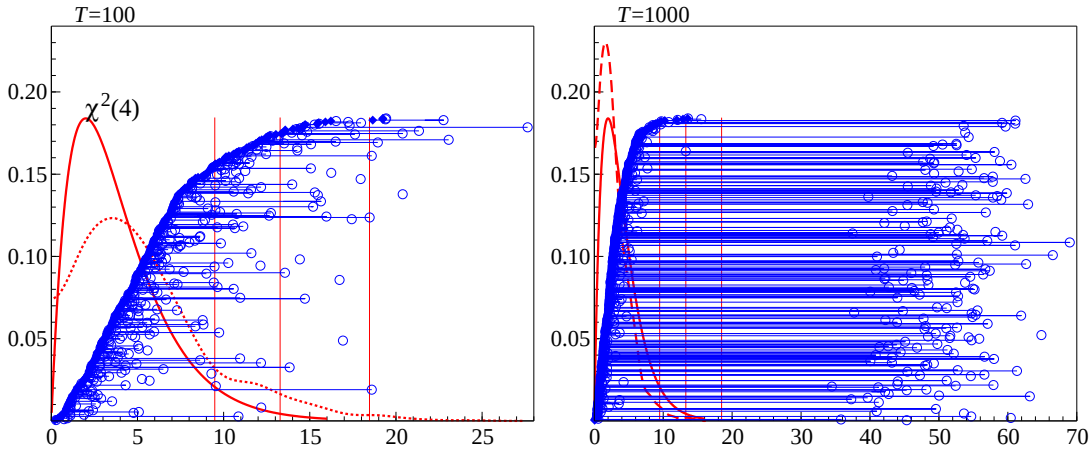

Figure 3. Formula I(2): I(2)-D, $p = 6$, $\omega = 0.9$, $\rho_1 = 0$ laps with distant convergence. (**Left**) $T = 100$, (**Right**) $T = 1000$. Three extreme outliers for $T = 1000$ have been removed for readability. See caption of Figure 1 for more details.

Table 5. Formula I(2). Performance for restrictions I(2)-A, ..., I(2)-E, for 1000 laps. '-' means exactly zero, the other figures are percentages rounded to two decimals and multiplied by 100. Empty cells mean that the team did not take part in the race.

Res	k	T	p	ρ_0, ρ_1	Team 1						Team 2						Team 3						DNF	NOR
					SC	WC	DC	FC	AD	IT	SC	WC	DC	FC	AD	IT	SC	WC	DC	FC	AD	IT		
A	2	100	6	0.0,0.0	100	–	–	–	–	3	100	–	–	–	–	3	100	–	–	–	–	40	–	1
A	2	100	6	0.0,0.9	100	–	–	–	–	5	100	–	–	–	–	5	100	–	–	–	–	49	–	1
A	2	100	6	0.9,0.0	100	–	0	–	2	11	100	–	0	–	2	14	100	–	–	–	–	71	–	1.00
A	2	100	6	0.9,0.9	99	–	1	–	1	14	99	–	1	0	1	17	100	0	–	–	–	75	–	1.01
A	2	1000	6	all	100	–	–	–	–	1	100	–	–	–	–	1	100	–	–	–	–	32	–	1
B	2	100	6	0.0,0.0	73	0	27	–	3	40	69	1	30	0	3	128	98	1	1	–	1	184	–	1.33
B	2	100	6	0.0,0.9	79	1	20	–	3	27	76	2	22	–	3	78	98	2	0	–	6	158	–	1.23
B	2	100	6	0.9,0.0	95	0	5	–	2	19	93	1	6	1	2	35	99	1	0	–	1	134	–	1.06
B	2	100	6	0.9,0.9	96	0	3	–	2	19	95	1	4	0	2	39	99	1	0	–	1	117	–	1.04
B	2	1000	6	0.0,0.0	71	3	26	0	4	100	67	6	27	1	3	184	89	5	6	–	2	252	–	1.41
B	2	1000	6	0.0,0.9	71	3	26	–	3	50	70	3	28	–	4	105	95	2	2	–	0	182	–	1.37
B	2	1000	6	0.9,0.0	74	2	24	–	10	56	71	2	26	1	10	135	97	2	1	–	3	195	–	1.30
B	2	1000	6	0.9,0.9	73	1	26	–	10	33	73	1	27	0	10	85	99	1	–	–	–	160	–	1.28
C	2	100	6	0.0,0.0							100	–	–	–	–	4	100	–	–	–	–	35	–	1
C	2	100	6	0.0,0.9							100	–	–	–	–	7	100	–	–	–	–	39	–	1
C	2	100	6	0.9,0.0							100	–	0	–	2	11	100	–	–	–	–	35	–	1.00
C	2	100	6	0.9,0.9							99	0	1	–	2	11	100	–	0	–	0	32	–	1.01
C	2	1000	6	all							100	–	–	–	–	1	100	–	–	–	–	30	–	1
D	2	100	6	0.0,0.0	74	4	12	10	8	212	61	1	4	34	12	136	90	4	6	0	0	416	–	1.21
D	2	100	6	0.0,0.9	77	4	17	2	19	163	59	2	14	26	14	210	73	8	19	0	6	514	–	1.43
D	2	100	6	0.9,0.0	76	5	12	7	1	174	72	1	3	24	1	134	90	4	5	0	0	390	–	1.17
D	2	100	6	0.9,0.9	72	4	21	3	1	141	59	3	13	25	1	189	79	9	12	–	1	433	–	1.38
D	2	1000	6	0.0,0.0	41	14	9	37	414	263	35	19	4	42	543	296	66	8	17	8	0	708	0.03	1.26
D	2	1000	6	0.0,0.9	66	8	18	7	25	253	48	4	17	32	30	318	63	8	26	3	5	683	0.00	1.54
D	2	1000	6	0.9,0.0	35	22	3	41	31	241	40	29	3	28	12	279	85	7	5	2	0	568	0.01	1.09
D	2	1000	6	0.9,0.9	67	8	18	7	14	230	49	5	15	31	14	258	79	6	14	0	1	522	–	1.44
E	2	100	6	0.0,0.0	46	3	42	10	2	193	36	2	38	24	3	284	79	3	15	3	1	427	0.01	1.76
E	2	100	6	0.0,0.9	61	2	35	2	4	116	57	2	27	14	4	156	85	4	9	2	1	302	0.00	1.59
E	2	100	6	0.9,0.0	75	4	15	6	2	147	64	2	12	22	2	159	84	6	9	1	1	409	0.00	1.27
E	2	100	6	0.9,0.9	73	3	22	1	2	102	61	3	22	14	2	169	81	6	13	–	1	374	–	1.44
E	2	1000	6	0.0,0.0	38	6	35	21	27	358	27	8	39	26	12	450	49	8	39	4	5	665	0.01	1.91
E	2	1000	6	0.0,0.9	49	4	41	7	7	247	33	4	47	16	6	353	52	4	41	3	3	511	0.01	2.11
E	2	1000	6	0.9,0.0	36	7	44	14	13	310	22	6	50	23	13	337	69	5	23	4	5	640	0.00	1.99
E	2	1000	6	0.9,0.9	58	3	37	2	10	130	44	5	43	9	10	235	69	5	26	1	6	374	0.00	1.84

As the final aspect of Formula I(2), consider the misspecified restrictions on τ in model I(2)-E. Figure 4 shows that distant convergence has no practical implication for large samples ($T = 1000$), where the power would be 1 anyway. Conversely, in small samples ($T = 100$) distant convergence slightly increases the rejection rate, which would be quite high in any case.

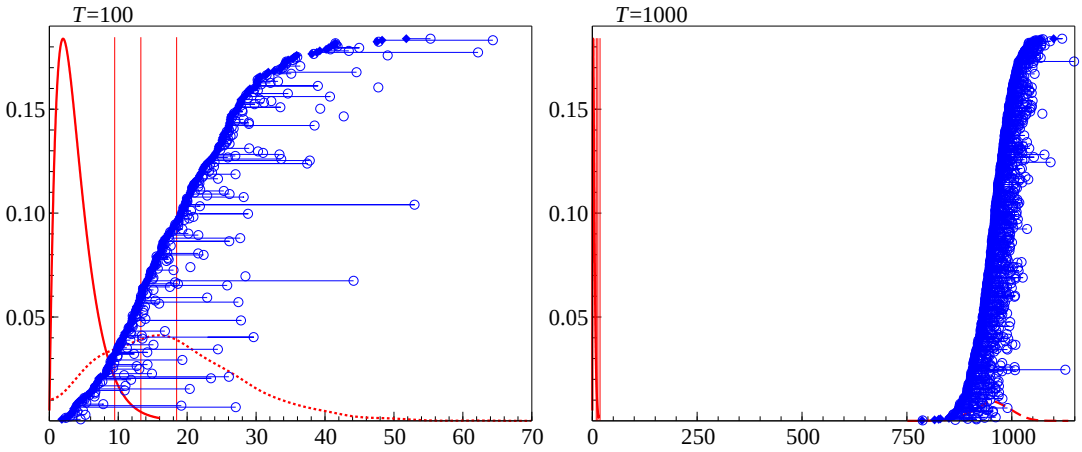

Figure 4. Formula I(2): I(2)-E, $p = 6$, $\omega = 0.9$, $\rho_1 = 0$ laps with distant convergence. (**Left**) $T = 100$, (**Right**) $T = 1000$. See caption of Figure 1 for more details.

Overall, in the setting of Formula I(2), maximizing the likelihood under correctly specified restrictions on α and β seems fast and accurate. Conversely, when correctly specified restrictions on τ are introduced, finding the overall maximum of the likelihood is not easy. Since β is one of the components of τ, one might guess that the problems arise from the complementary directions with respect to β within τ; the issue deserves further exploration.

As in the I(1) case, maximizing the likelihood under misspecified restrictions is difficult; however, the consequence of this difficulty are benign, because they appear to increase the power of the test for the current design of the Formula I(2) races.

9. Conclusions

The test run of the championships shows that there is room for improving algorithms. It demonstrates the strength of this 'collective learning' experiment, where other researchers may try and propose new algorithm to improve on the existing ones. All algorithms win in the end, since each team learns where and how to improve the algorithm design.

Other circuits may be added in the future, as algorithms improve. Races with a similar spirit can be set up in other adjacent fields, like fractional cointegration; the same principles may in fact be applied to any other model classes where maximizing the likelihood needs numerical optimization.

Acknowledgments: Financial support from the Robertson Foundation (Award 9907422) and the Institute for New Economic Thinking (Grant 20029822) is gratefully acknowledged by the first author and from the Italian Ministry of Education, University and Research (MIUR) PRIN Research Project 2010–2011, prot. 2010J3LZEN, 'Forecasting economic and financial time series' by the second author. The authors thank two anonymous referees for useful comments on the first version of the paper.

Author Contributions: The authors contributed equally to the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Practical Requirements for Submission

To facilitate submission and automated processing of results, some conventions are established that submissions to the project must adopt.

Appendix A.1. Innovations

A file containing 12 000 i.i.d. $N(0,1)$ time series of length 1000 is provided (`ERRORS.CSV`) in the companion website. The series are organized column-wise and labelled `eps00001` to `eps12000`. In other words, this file contains the $1000 \times 12\,000$ matrix E. The p-dimensional vector $\varepsilon_t^{(i)}$, $t = 1, ..., T$, is obtained as the transpose of the t-th row of the submatrix $E(1 : T, [i-1]p + 1 : ip)$, assuming indexation starts at element $(1, 1)$.

Table A1 provides the first five generated observations for lap 1 of the I(1) and I(2) DGP with $p = 6, \rho_0 = \rho_1 = \omega = 0.9$.

Table A1. The first five observations of the generated data for I(1) and I(2) DGPs with $p = 6, \rho_0 = \rho_1 = \omega = 0.9$. Ten significant digits given; computation uses double precision.

t			**Formula I(1)** $X_t^{(1)\prime}$			
1	0.2548828200	−2.009603960	0.5542620800	0.7913726500	−0.5458015100	−1.349741980
2	0.7806863280	−6.446826254	0.1020649020	−1.468146855	−1.017498239	−2.539647722
3	−0.3490545448	−9.526135279	−0.08454244820	−1.010888930	0.04508386490	−0.3954565398
4	−0.4230090503	−11.98920553	−0.6228956034	−2.179696857	−1.063624342	−1.528447976
5	−0.09820491529	−14.61563411	−1.559382683	−1.481900221	0.05204249257	−0.4856795382
t			**Formula I(2)** $X_t^{(1)\prime}$			
1	0.2548828200	−2.009603960	0.5542620800	0.7913726500	−0.5458015100	−1.349741980
2	0.8061746100	−6.647786650	0.1020649020	−0.6767742050	−0.7626154190	−4.549251682
3	−0.2454976300	−10.37177830	−0.08454244820	−1.687663135	0.8257701929	−6.842282794
4	−0.3543575900	−13.78746208	−0.6228956034	−3.867359991	−1.412678886	−11.05458325
5	−0.07185436000	−17.61281121	−1.559382683	−5.349260212	−0.3709665578	−12.47488507

Appendix A.2. Report File Naming

For each circuit, a team needs to upload an output file on the companion website with either `txt` or `csv` extension. The former is a text file where numbers are separated by a space, while the latter is a csv spreadsheet file using a comma as separator (and without column headers). In all cases there will be one lap per line in the output file.

The output file should be named `FIxDGPyyyMODzzz.csv` (or `FIxDGPyyyMODzzz.txt`), where:

x 1 for Formula I(1), 2 for Formula I(2);
yyy three digits DGP index n, as defined in Table A2;

Table A2. Definition of the DGP index n.

DGP index $n := 8i_T + 4i_p + 2i_0 + i_1 + 1$			
$i_T = 0$	$T = 100$	$i_0 = 0$	$\rho_0 = 0$ for Formula I(1) or $\omega = 0$ for Formula I(2)
$i_T = 1$	$T = 1000$	$i_0 = 1$	$\rho_0 = 0.9$ for Formula I(1) or $\omega = 0.9$ for Formula I(2)
$i_p = 0$	$p = 6$	$i_1 = 0$	$\rho_1 = 0$
$i_p = 1$	$p = 12$	$i_1 = 1$	$\rho_1 = 0.9$

zzz three digits model index m, as defined in Table A3;

Table A3. Definition of the model index m.

Model index $m := 2i_r + i_k + 1$	
$i_r = 0$	Restriction I(1)-A or I(2)-A
$i_r = 1$	Restriction I(1)-B or I(2)-B
$i_r = 2$	Restriction I(1)-C or I(2)-C
$i_r = 3$	Restriction I(2)-D
$i_r = 4$	Restriction I(2)-E
$i_r = 4 + r + (r+s-1)(r+s)/2$	$\mathcal{M}(r,s)$ with ordering as in Table A4
$i_k = 0$	$k = 2$
$i_k = 1$	$k = 5$

The ordering of the unrestricted I(2) estimates corresponds to the column vectorization of the upper diagonal of the relevant part of a ranks test table. For instance, in case $p = 6$ the ordering of models is the one in Table A4.

Table A4. Ordering of the models $\mathcal{M}(r,s)$ for case $p = 6$. Entries in the table correspond to the numbering of models, where $s_2 = p - r - s$. The ordering is similar for the case $p = 12$.

$r \backslash s_2$	**5**	**4**	**3**	**2**	**1**
1	1	2	4	7	11
2		3	5	8	12
3			6	9	13
4				10	14
5					15

As an example, results for the Formula I(2) circuit with $n = 13$ ($i_T = 1$, $i_p = 1$, $i_0 = 0$ and $i_1 = 0$) and $m = 6$ ($i_r = 2$, $i_k = 1$), should be stored in a file named `FI2DGP013MOD006.csv` (or `FI2DGP013MOD006.txt`).

Appendix A.3. Report File Content

Formula I(1) files have N lines with $4 + 2\,(p+1)\,r$ numbers, whereas Formula I(2) files have N lines, each with $4 + 2\,(p+1)\,r + p\,(p+1)$ numbers. Each line contains the following information:

$$(i : \ell^u_{a,c,i} : \ell_{a,c,i} : N_{a,c,i} : S_{a,c,i} : \theta^{R\prime}_{a,c,i}),$$

where:

1. i is the lap number, $i = 1, ..., 1000$;
2. $\ell^u_{a,c,i}$ is the unrestricted loglikelihood, reported with at least 8 significant digits:
   - Formula I(1): loglikelihood of the unrestricted I(1) model;
   - Formula I(2) $i_r > 4$: loglikelihood of the VAR;
   - Formula I(2) $i_r \leq 4$: loglikelihood of the unrestricted I(2) model.
3. $\ell_{a,c,i}$ as defined in (2) with at least 8 significant digits;
4. $N_{a,c,i}$, the iteration count;
5. $S_{a,c,i}$ is the integer convergence indicator, 1 for convergence, 0 for no convergence;
6. $\theta^{R\prime}_{a,c,i}$ is part of the coefficient vector, which is for Formula I(1):

$$\theta^{R\prime}_{a,c,i} = \left(\mathrm{vec}(\alpha_{a,c,i})' : \mathrm{vec}(\beta_{a,c,i})'\right).$$

For the Formula I(2) circuits use instead:

$$\theta^{R\prime}_{a,c,i} = \left(\mathrm{vec}(\alpha_{a,c,i})' : \mathrm{vec}(\beta_{a,c,i})' : \mathrm{vec}(\Gamma : \mu_0)'_{a,c,i}\right).$$

Coefficients must be reported exactly in the given order, providing at least 8 significant digits (but 15 digits is recommended). No particular normalization is required.

If the algorithm failed because likelihood evaluation failed (e.g., singular Ω), then $\ell_{a,c,i} = -\infty$ should be reported. The data is processed with Ox, so `.NaN` and `.Inf` are allowed. Because there is no clear convention on writing $-\infty$, any value of -10^{308} or lower is interpreted as $-\infty$.

Table A5 provides the start of the first three lines of three selected output files.

Table A5. Three examples of output files. Beginning of first three lines given.

```
                                FI1DGP001MOD001.csv
1,      84.7587177451401,      82.2423190842343,      3, 1,-0.187577914295476,  ...
2,      30.1177483889851,      28.6953188342152,      5, 1,0.299447436254108,   ...
3,      64.5916602781794,      59.9799720330047,      4, 1,-0.0746786280148741, ...

                                FI1DGP001MOD001.txt
1    84.75871775    82.24231908    10    1 -1.8757787e-001  1.6004096e-002 ...
2    30.11774839    28.69531883    20    1  2.9944720e-001 -5.8528461e-002 ...
3    64.59166028    59.97997203    16    1 -7.4678444e-002 -1.5937408e-001 ...

                                FI2DGP001MOD001.csv
1,      76.4430824288192,      76.2400219176979,      10, 1,0.0844284641160844, ...
2,      27.5347594941493,      26.6849814069451,      19, 1,0.0711585542069055, ...
3,       48.709883495749,      48.2827756129209,      24, 1,0.12242477122602,   ...
```

References

Abadir, Karim M., and Paolo Paruolo. 2009. On efficient simulations in dynamic models. In *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry.* Edited by Jennifer Castle and Neil Shephard. Oxford: University Press, pp. 270–301.

Anderson, Theodore Wilbur. 1951. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics* 22: 327–51. Correction in Annals of Statistics 8, 1980: 1400.

Beiranvand, Vahid, Warren Hare, and Yves Lucet. 2017. Best practices for comparing optimization algorithms. *Optimization and Engineering* 18: 1–34.

Boettiger, Carl. 2015. An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review* 49: 71–79.

Boswijk, H. Peter. 2000. Mixed normality and ancillarity in I(2) systems. *Econometric Theory* 16: 878–904.

Boswijk, H. Peter, and Jurgen A. Doornik. 2004. Identifying, estimating and testing restricted cointegrated systems: An overview. *Statistica Neerlandica* 58: 440–65.

Boswijk, H. Peter, and Paolo Paruolo. 2017. Likelihood ratio tests of restrictions on common trends loading matrices in I(2) VAR systems. *Econometrics* 5: 28.

Doornik, Jurgen A. 1998. Approximations to the asymptotic distribution of cointegration tests. *Journal of Economic Surveys* 12: 573–93. Reprinted in Michael McAleer and Les Oxley. 1999. *Practical Issues in Cointegration Analysis*. Oxford: Blackwell Publishers.

Doornik, Jurgen A. 2013. *Object-Oriented Matrix Programming Using Ox*, 7th ed. London: Timberlake Consultants Press.

Doornik, Jurgen A. 2017a. *Accelerated Estimation of Switching Algorithms: The Cointegrated VAR Model and Other Applications*. Working Paper 2017-W05. Oxford: Nuffield College.

Doornik, Jurgen A. 2017b. Maximum likelihood estimation of the I(2) model under linear restrictions. *Econometrics* 5: 19.

Doornik, Jurgen A., and David F. Hendry. 2013. *Modelling Dynamic Systems Using PcGive: Volume II*, 5th ed. London: Timberlake Consultants Press.

Hendry, David F. 1984. Monte Carlo experimentation in econometrics. In *Handbook of Econometrics*. Edited by Zvi Griliches and Michael D. Intriligator. New York: North-Holland, vol. 2, pp. 937–76.

Johansen, Søren. 1988. Statistical Analysis of Cointegration Vectors. *Journal of Economic Dynamics and Control* 12: 231–54.

Johansen, Søren. 1991. Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica* 59: 1551–80.

Johansen, Søren. 1995a. Identifying restrictions of linear equations with applications to simultaneous equations and cointegration. *Journal of Econometrics* 69: 111–32.

Johansen, Søren. 1995b. A statistical analysis of cointegration for I(2) variables. *Econometric Theory* 11: 25–59.

Johansen, Søren. 1997. A likelihood analysis of the I(2) model. *Scandinavian Journal of Statistics* 24: 433–62.

Johansen, Søren, and Katarina Juselius. 1992. Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for UK. *Journal of Econometrics* 53: 211–44.

Johansen, Søren, and Katarina Juselius. 1994. Identification of the long-run and short-run structure: an application of the ISLM model. *Journal of Econometrics* 63: 7–36.

Mosconi, Rocco, and Paolo Paruolo. 2016. *Cointegration and Error Correction in I(2) Vector Autoregressive Models: Identification, Estimation and Testing*. Mimeo: Politecnico di Milano.

Mosconi, Rocco, and Paolo Paruolo. 2017. Identification conditions in simultaneous systems of cointegrating equations with integrated variables of higher order. *Journal of Econometrics* 198: 271–76.

Noack Jensen, Anders. 2014. Some Mathematical and Computational Results for Vector Error Correction Models. Chapter 1: The Nesting Structure of the Cointegrated Vector Autoregressive Model. Ph.D. Thesis, University of Copenhagen, Department of Economics, Copenhagen.

Onatski, Alexei, and Harald Uhlig. 2012. Unit roots in white noise. *Econometric Theory* 28: 485–508.

Paruolo, Paolo. 2002. On Monte Carlo estimation of relative power. *Econometrics Journal* 5: 65–75.

Paruolo, Paolo. 2005. *Design of Vector Autoregressive Processes for Invariant Statistics*. WP 2005-6. Insubria: University of Insubria, Department of Economics.

Paruolo, Paolo, and Anders Rahbek. 1999. Weak exogeneity in I(2) VAR systems. *Journal of Econometrics* 93: 281–308.

Rahbek, Anders C., Hans Christian Kongsted, and Clara Jørgensen. 1999. Trend-Stationarity in the I(2) Cointegration Model. *Journal of Econometrics* 90: 265–289.