**Revista Colombiana de Estadística - Applied Statistics**
January 2025, volume 48, issue 1, pp. 71–98
http://doi.org/10.15446/rce.v48n1.113209

# Nowcasting the State of the Economy: An Application of Linear Combinations of Dynamic Common Factors to the Colombian Economy

## Prediciendo el estado de la economía: una aplicación de combinaciones lineales de factores comunes dinámicos a la economía colombiana

Mario Arrieta-Prieto[a], Fabio H. Nieto[b]

Department of Statistics, School of Sciences, Universidad Nacional de Colombia, Bogotá, Colombia

**Abstract**

The main goal of this work is to propose a general methodology to create a coincident index for the economy of a given country or region based on linear combinations of dynamic common factors, and to validate it on simulated scenarios and on a case study for Colombia. The methodology proposed can effectively handle both stationary and nonstationary macroeconomic variables as input and provides tools to obtain point estimates and confidence regions, and to test hypotheses for the linear-combination coefficients. This work highlights how promising this new proposal is in terms of its contribution with respect to its antecedents in the literature, by showcasing situations in which a linear combination of dynamic factors can enhance the accuracy of the nowcast and address potential problems and limitations of considering only one dynamic factor as index. The application of this work to the Colombian economy is based on the macroeconomic analysis conducted by previous researchers. This work, however, considers and analyzes a much larger set of candidate indices via linear combinations of factors, providing consistent results; which strengthens and validates their previous findings.

***Key words*:** Coincident index; Genetic algorithms; Macroeconomics; Nonconvex optimization; Multivariate time series.

[a]Ph.D. E-mail: mearrietap@unal.edu.co
[b]Ph.D. E-mail: fhnietos@unal.edu.co

**Resumen**

El objetivo principal de este trabajo es proponer una metodología general para crear un índice coincidente para la economía de un país o región determinado, basada en combinaciones lineales de factores comunes dinámicos, y validarla en escenarios simulados y en un estudio de caso para Colombia. La metodología propuesta puede manejar de manera efectiva tanto variables macroeconómicas estacionarias como no estacionarias como datos de entrada y proporciona herramientas para obtener estimaciones puntuales y regiones de confianza, así como para probar hipótesis sobre los coeficientes de la combinación lineal. Este trabajo destaca lo prometedora que es esta nueva propuesta en términos de su contribución con respecto a sus antecedentes en la literatura, al mostrar situaciones en las que una combinación lineal de factores dinámicos puede mejorar la precisión del pronóstico del estado de la economía, además de abordar problemas y limitaciones potenciales al considerar solo un factor dinámico como índice. La aplicación de este trabajo a la economía colombiana se basa en el análisis macroeconómico realizado por investigadores previos. Sin embargo, este trabajo considera y analiza un conjunto mucho más amplio de índices candidatos a través de combinaciones lineales de factores, proporcionando resultados consistentes, lo que fortalece y valida sus hallazgos anteriores.

***Palabras clave*:** Índice coincidente; Algoritmos genéticos; Macroeconomía; Optimización no convexa; Series de tiempo multivariadas.

# 1. Introduction

Economic indices are of vital importance for the macroeconomic planning of a given country. Particularly, coincident indices attempt to predict the state of the economy in a given time point, based on the available information up to that point. It is well known that the Gross Domestic Product (GDP) is a massive undertaking to measure the overall performance of the economy; however, it is not calculated at a high frequency because it is extremely time-consuming. For that reason, simpler alternative coincident indices are necessary.

According to Stock & Watson (1992), an economic index corresponds to the estimation, or the prediction, of the realization for a non-observable variable: the state of the economy for a specific country or region, based on the information taken from a set of observable macroeconomic variables denominated indicators of the economy.

Even if each one of the indicator variables for the economy can show erratic behaviors and different temporal trajectories from each other, there is a latent influence of the state of the economy over all the variables, providing them common characteristics. The essence in the construction of an economic index, generally speaking, lies on the correct identification and careful isolation of the common information in the set of indicator variables.

Particularly, a coincident index for the state of the economy must be able to predict the state of the economy for a given instant of time by using the available information up to that time instant, i.e., it must match the *business cycle* of the

economy, which is defined as the representing cycle of the characteristic oscillations of the macroeconomic activity (Burns & Mitchell, 1946). Altissimo et al. (2010) define as the main objective for a coincident index to do a valuation about the state of the economy that is (1) *comprehensive and non-subjective*, which means that it has to condense in a proper way the information of the indicator variables with no place for subjectivity biases; (2) *timely*, since it has to provide real-time estimates using all the information at hand; and (3) *free from short-run fluctuations*, to capture and show the actual trend for the state of the economy without any transient perturbations.

The first attempts to create an economic index were based on heuristics. An economic index was computed as a weighted average of the indicator variables in such a way that the weights for each variable were assigned based on criteria and general knowledge of the context, but with no statistical foundation (Martínez et al., 2016). Stock & Watson (1992) were the first ones who proposed the idea of the state of the economy as a latent stochastic process that is related in a linear way with each of the indicator variables. They formulated their model in terms of the first differences of both the observed and the latent variables involved, to ensure stationarity. Further developments in the area started considering the concept of Dynamic Common Factors (DCFs) as an alternative to capture different and multiple common trends affecting the macroeconomic indicators. This approach generalized the idea of only one common factor proposed by Stock & Watson (1992). Following the methodology proposed by Martínez et al. (2016) and Chudt & Nieto (2018), this work proposes a method to optimally combine the estimated factors for a set of macroeconomic variables in a linear fashion, to provide a broader spectrum of possibilities to be considered as coincident indices. Besides, general inference procedures are developed to compare and establish the added value of the proposed methodology with respect to previous developments in both simulated scenarios and a case study for the Colombian economy.

The rest of this manuscript is as follows. Section 2 provides the general notation and mathematical prerequisites to understand the methodological innovation proposed. Section 3 describes the methodology to optimally estimate the coefficients of a lineal combination of estimated factors that could play the role of a coincident index, Section 4 describes the procedures to conduct interval estimation and hypothesis testing on those coefficients, and Section 5 presents a simulation procedure to exhaustively test the methodology proposed. Finally, Section 6 presents the results of the simulated instances and the application to the data gathered for Colombia (South America) while Section 7 summarizes the conclusions of this work.

## 2. DCF-based Coincident Index: A Review

Let $\{Y_t\}$ be a multivariate stochastic process of dimension $m$, and let $\{f_t\}$ be a multivariate latent stochastic process of dimension $r$, $r < m$, that is related to $\{Y_t\}$ via the equation

$$Y_t = Pf_t + e_t, t \in \mathbb{Z}, \tag{1}$$

where $\left\{ f_t = (f_{1t}, f_{2t}, \ldots, f_{rt})^T \right\}$ is denominated the vector of DCFs of the process $\{Y_t\}$, $^T$ is the transpose operator, $P \in \mathbb{R}^{m \times r}$ is a matrix of weights of the DCFs and $\{e_t\}$ is a (Gaussian) noise process of dimension $m$, whose variance-covariance matrix is denoted by $\Sigma_e$. In the economic theory, the realizations of $\{Y_t\}$ are the set of macroeconomic variables or indicators of the economic activity. The realization of the process $\{f_t\}$ carries all the common characteristics of the macroeconomic indicators in a lower-dimension object. The main goal of creating an economic index is, then, to extract as much information of the macroeconomic indicators $\{Y_t\}$ in a compact way by means of $\{f_t\}$; and, based on the factors, to compute another process denominated *index*, $\{I_t\}$, that accurately resembles the behavior of the latent stochastic process $\{C_t\}$, the so-called state of the economy or reference cycle.

According to Wei, William WS (2006) and confirmed by Martínez et al. (2016), when dealing with nonstationary vector time series, differencing allows to achieve stationarity but might eliminate and distort some of the interrelationships that the series naturally have. For that reason, the methodology followed in this work considers approaches that allow the macroeconomic indicators to be realizations of nonstationary processes and handle this characteristic in an effective way.

Following this principle, Martínez et al. (2016) propose a four-step methodology, based on DCFs, that selects one of the estimated factors (from potentially nonstationary and cointegrated time series) as a coincident index. Reconsider the model expressed in Equation (1) assuming that $\{f_t\}$ follows a $VARMA\,(p, q)$ model that looks like

$$\Phi\,(B)\, f_t = \kappa + \Theta\,(B)\, a_t, \tag{2}$$

where the operator $\Phi\,(B)$ is such that all the roots of the complex polynomial $|\Phi\,(z)|$ lie outside or on the unit circle ($|\cdot|$ represents the determinant operation). Additionally, the operator $\Theta\,(B)$ is such that the complex polynomial $|\Theta\,(z)|$ has all its roots outside of the unit circle and distinct from the ones of $|\Phi\,(z)|$. The process $\{a_t\}$ corresponds to a multivariate Gaussian (denoted as $MVN$ later on) white noise process, independent of the Gaussian white noise process $\{e_t\}$; and whose variance-covariance matrix is of full rank and denoted as $\Sigma_a$.

Equations (1) and (2) facilitate to build the state-space representation of the dynamic factor model, provided that the model is identifiable. Their parameters can be estimated using Gaussian Maximum Likelihood or Expectation-Maximization algorithms, and the estimates of the factors can be obtained by means of the fixed-point smoother, which is based on the Kalman Filter. Lütkepohl (2005) points out that the Gaussian Maximum Likelihood and the Kalman Filter are reasonable approaches even when there is no normality in the white noise processes. It is important to mention that the model presented in Equation (1) suffers from identifiability issues that impede any procedure from retrieving a unique estimate of $P$ and the DCFs $\{f_t\}$, ergo, some identifiability restrictions need to be imposed. It is said that estimates of the factors are identifiable up to an orthogonal rotation (Peña & Tsay, 2021).

The four-step methodology proposed by Martínez et al. (2016) to design a coincident index is as follows:

1. *Adaptation and preparation of the time series.* First of all, it is necessary to deseasonalize the macroeconomic time series, and to pre-whiten the process $\{\hat{e}_t\}$, if it does not exhibit white noise features after a first round of estimation. In a parallel effort, Nieto et al. (2016) propose an alternative methodology to deal with seasonality in DCF models.

2. *Estimation of the common factors.* This step of the modeling is conducted based on the results by Peña & Poncela (2006). Based on the sample generalized covariance matrices (SGCV)

$$C_Y(k) = \frac{1}{S^{2d+1}} \sum_{t=k+1}^{S} \left(Y_{t-k} - \bar{Y}\right)\left(Y_t - \bar{Y}\right)^T,$$

and the canonical correlation matrices (CCM)

$$\hat{M}_1(k) = \left[\sum_{t=k+1}^{S} Y_t Y_t^T\right]^{-1} \left[\sum_{t=k+1}^{S} Y_t Y_{t-k}^T\right] \left[\sum_{t=k+1}^{S} Y_{t-k} Y_{t-k}^T\right]^{-1} \left[\sum_{t=k+1}^{S} Y_{t-k} Y_t^T\right],$$

where $S$ is the sample size, $d$ is the order of integration for the vector $\{Y_t\}$ and $k(= 0, 1, 2, \ldots)$ is a particular lag; they created a test to identify the number of common factors and a procedure to find their nature (if they are stationary or not). The test for the number of factors is based on an asymptotic result and its limiting distribution is independent of the lag $k$ considered, even if its expression does depend on it. However, when applied in finite-sample scenarios, the test is sensitive to the lag $k$ considered and the conclusions may vary according to the value it takes for a given confidence level due to sample variability. For that reason, some caution must be exercised when using this test for small sample sizes and sensitivity to the specification of the lag $k$ should be explored.

   Bujosa et al. (2013) propose an alternative approach based on the SCGV when the Gaussian, zero-mean and full rank variance-covariance matrix white noise assumptions do not hold.

3. *Choice of a common factor as the coincident index.* Taking as a reference the seminal work of Banerji (1999) to identify a leading index using the concept of a leading profile, Martínez et al. (2016) define a coincident profile as a tool to assess the adequacy of each estimated factor to be a coincident index with respect to a proxy for the state of the economy. The coincident profile is a synthesized presentation of several $p$-values that are the result of a nonparametric test of $h$-coincidence, with $h \in \mathbb{Z}$ representing the number of periods in advance ($h \geq 0$) or delayed ($h < 0$) for the factor to exhibit the same behavior (pattern) in terms of peaks, valleys, growth and decay periods of the state of the economy. The general hypothesis system of an $h-$coincidence test has the plausibility of $h-$coincidence, for a given value of $h$, in the null hypothesis;

against an alternative rejecting that value of $h$. Thus, leading indices should have high $p$-values for positive values of $h$ (indicating that the null hypothesis cannot be rejected), coincident indices should have high $p$-values for $h = 0$, and lagged indices should have high $p$-values for $h < 0$. For a formal definition of the $h-$coincidence test and its use for identifying coincident indices, see Banerji (1999), Martínez et al. (2016), Chudt & Nieto (2018).

As it has been mentioned before, the state of the economy, $\{C_t\}$, is a latent stochastic process; therefore, its realizations are not observable. For that reason, it is necessary to identify a good proxy, $\{\hat{c}_t\}$, of the state of the economy, to compare with when constructing the coincident profile. In most of the cases, a high-frequency interpolated series of the GDP is a reasonable candidate for this task. Then, the dynamic factor with the highest acceptable $p$-value of 0-coincidence (above a certain significance level) is chosen as a coincident index, implying that the null hypothesis of 0-coincidence between that factor and the proxy cannot be rejected. At this point, it must be clear that the proxy, $\{\hat{c}_t\}$, acts as a coincident index itself; however, a massive effort is usually required to compute it, so the main idea is to identify an equally reliable coincident index that is easier to compute by using the estimated factors of some macroeconomic series.

Chudt & Nieto (2018) mention that the coincident profile procedure sometimes does not work with the original proxy and the macroeconomic variables because the series can be excessively smooth (lacking of turning points) or excessively noisy. Therefore, they applied the coincident profile procedure to the first differences of the proxy and each of the factors, leveraging the general fact that if any two continuous functions coincide in their rates of change over time (have the same derivative, if seen as continuous functions of time), they are identical up to an additive constant.

4. *Identification of the basis for the index.* Once one of the factors has been identified as a coincident index, it is necessary to establish the temporal basis for which the selected factor behaves as a coincident index and analyze its implications in the given context.

Martínez et al. (2016) also pose for future work the possibility of constructing a coincident index as a linear combination of the DCFs and assessing its statistical adequacy to predict the state of the economy, which is the main goal of this work.

## 3. Methodology

Previous efforts in this area have focused their attention on selecting only one among the estimated factors as the coincident index based on their coincident profiles. Nevertheless, this approach is somehow restrictive because it does not exploit the possible synergies that might exist between the factors that could lead to a better index in terms of similarity with the proxy for the state of the economy.

The coincident profile proposed by Martínez et al. (2016), and then refined by Chudt & Nieto (2018), is the presentation of the $p$-value for several tests of $h$-coincidence between the separate components of $\left\{\hat{f}_t\right\}$ (the estimated factors) and $\{\hat{c}_t\}$ (the proxy of the state of the economy), with $h \in \{-3, -2, -1, 0, 1, 2, 3\}$. In other words, Martínez et al. (2016) and Chudt & Nieto (2018) considered only trivial combinations of the factors. For each one of the estimated factors, the level of $h$-coincidence is determined by choosing the value of $h$ for which the $p$-value for $h$-coincidence is the maximum among the coincident profile, over a pre-specified significance level. For coincident indices, the attention focuses on the case when $h = 0$.

Since the search space of candidate factors to be a coincident index is finite and very reduced under this approach, it is not always possible to find a factor that is 0-coincident. This limitation led the researchers to choose sometimes as coincident index a factor with a coincidence level other than 0-coincidence (but fairly close) if 0-coincidence is not achieved.

Contrastingly, this work intends to explore all the possible linear combinations of estimated factors to identify a particular combination that maximizes 0-coincidence (considering other elements that will be introduced later).

In essence, the idea is to create an economic index of the form

$$I_t(\vec{\alpha}) := \vec{\alpha}^T f_t = \sum_{j=1}^{r} \alpha_j f_{jt}, t \in \mathbb{Z},$$

being $\alpha_j \in \mathbb{R}, j = 1, 2, \ldots, r$. Its data-based counterpart, computed with the estimated factors, $\left\{\hat{f}_t\right\}$; will be denoted as $\{i_t\}$.

Given that the 0-coincidence is based on a statistical test that assesses the coincidence between the turning points of two series (Banerji, 1999), it is invariant under transformations of scale, thus any nonnegatively-scaled version of the candidate index will produce the same 0-coincidence. In other words, for any $k \in \mathbb{R}_+$ (with $\mathbb{R}_+$ the set of real positive numbers), $\left\{\vec{\alpha}^T f_t\right\}$ and $\left\{k\vec{\alpha}^T f_t\right\}$ have the same level of 0-coincidence with a given proxy.

For that reason, to ensure identifiability, the coefficients for the factors have to be normalized. In this work, the normalization is made on the basis of the $L_2$ norm, i.e.,

$$\|\vec{\alpha}\|_2^2 = \sum_{j=1}^{r} \alpha_j^2 = 1. \tag{3}$$

The feasible region is defined in terms of an equality and not as $\|\vec{\alpha}\|_2^2 \leq 1$ because, even if the latter is convex, it does not solve the identifiability problem since for any $\vec{\alpha}_0$ belonging to the feasible region, $k\vec{\alpha}_0$ would belong to it as well, for every $k \in (0, 1)$.

Besides, the $L_2$ norm was preferred over the $L_1$ norm, for instance, because $L_1$ usually is employed for variable (or factor, in this case) selection in regularization

models, which is undesirable if the idea is to look at nontrivial linear combinations of the factors.

A first attempt to address the problem of finding a coincident index by means of linear combinations of the DCFs would be to solve the following maximization problem

$$\begin{aligned} &\max_{\vec{\alpha}\in\mathbb{R}^r} p\left(\{\Delta\hat{c}_t\}, \{\Delta i_t\left(\vec{\alpha}\right)\}\right), \\ \text{s.t. } &i_t\left(\vec{\alpha}\right) = \sum_{j=1}^{r} \alpha_j \hat{f}_{jt}, \\ &\|\vec{\alpha}\|_2^2 = 1, \end{aligned} \tag{4}$$

being $\{\hat{c}_t\}$ a realization of the proxy for the state of the economy, $\{i_t\}$ a candidate linear combination of the estimated factors, $\Delta$ the finite difference operator and $p\left(\cdot,\cdot\right)$ the function that calculates the $p$-value for the 0-coincidence between two given series. The acronym "*s.t.*" stands for "*subject to:*" and precedes the constraints in the optimization problem. The difference operator is introduced following Chudt & Nieto (2018), as a way to effectively measure 0-coincidence by avoiding the excess of noise or smoothness that the proxy and the factors might exhibit. Nevertheless, the candidate index is still a function of both stationary and nonstationary factors, estimated to pontentially benefit from the cointegration relationships that the original macroeconomic variables might have.

Weierstrass' theorem guarantees that if the feasible region of an optimization problem is a compact set (which indeed is in this case) and the objective function is continuous, an optimal value exists within the feasible region (Bazaraa et al., 2013). However, as it is described in Subsection 3.2, the function $p\left(\cdot,\cdot\right)$ has multiple optima and that does not allow to obtain a unique solution. It is also important to highlight that there is no closed-form expression to evaluate $p\left(\cdot,\cdot\right)$ given any two arguments. This function has to be evaluated by means of a permutation test (Chudt & Nieto, 2018), leaving out of consideration any derivative-based optimization algorithm.

The following subsections present some inconveniences that impeded to tackle the optimization problem in Equation (4) directly and the remedial mechanisms proposed.

## 3.1. Use of Spherical Coordinates

The next step in the process of estimation for the linear-combination coefficients is to identify an algorithm to solve the optimization problem formulated in Equation (4). In addition to the inconveniences of multiple optima and the absence of derivatives, the $L_2$-norm constraint imposed to the coefficients for the factors in Equation (3) defines a non-convex space in $\mathbb{R}^r$, which makes it almost intractable for the iterative optimization algorithms usually implemented.

To overcome this challenge, the problem was solved using a spherical coordinate representation because it makes the feasible set convex. Recall that

for a point $(\alpha_1, \alpha_2, \ldots, \alpha_r) \in \mathbb{R}^r$, represented in the Cartesian coordinates system, there is an equivalent representation in the spherical coordinates system: $(\rho, \theta, \phi_1, \phi_2, \ldots, \phi_{r-2})^T$ where $\rho \in \mathbb{R}_+ \cup \{0\}$, $\theta \in [0, 2\pi]$ and $\phi_j \in [0, \pi], j \in \{1, 2, \ldots, r-2\}$. The equations that describe the relationship between these two representations, according to Blumenson (1960), are

$$\begin{aligned}
\alpha_r &= \rho \cos(\phi_{r-2}), \\
\alpha_{r-1} &= \rho \sin(\phi_{r-2}) \cos(\phi_{r-3}), \\
\alpha_{r-2} &= \rho \sin(\phi_{r-2}) \sin(\phi_{r-3}) \cos(\phi_{r-4}), \\
&\vdots \\
\alpha_2 &= \rho \sin(\phi_{r-2}) \sin(\phi_{r-3}) \sin(\phi_{r-4}) \sin(\phi_{r-5}) \ldots \sin(\phi_1) \sin(\theta), \\
\alpha_1 &= \rho \sin(\phi_{r-2}) \sin(\phi_{r-3}) \sin(\phi_{r-4}) \sin(\phi_{r-5}) \ldots \sin(\phi_1) \cos(\theta).
\end{aligned} \tag{5}$$

The constraint in Equation (3) is equivalent to the surface of a hypersphere in $\mathbb{R}^r$, therefore, its representation in spherical coordinates is simply

$$\begin{aligned}
&\rho = 1; \\
&0 \leq \phi_l \leq \pi, l \in \{1, 2, \ldots, r-2\}; \\
&0 \leq \theta \leq 2\pi.
\end{aligned} \tag{6}$$

The feasible region in Equation (6) corresponds to a polyhedron, which is a convex region. It is also important to notice that since $\rho = 1$, the feasible region in spherical coordinates can be seen as a subset of $\mathbb{R}^{r-1}$ whose variables are $(\theta, \phi_1, \phi_2, \ldots, \phi_{r-2})$. Once the optimal solution for this problem is computed, the estimates of the original coefficients, $\vec{\alpha}$, can be calculated by replacing the optimal values in spherical coordinates into the set of equations in Equation (5) with $\rho = 1$.

## 3.2. Inconveniences with 0-coincidence (Multiple Optima)

As mentioned before, the function $p(\cdot, \cdot)$ may have multiple optima (it might reach the value of 1 for different arguments). Figure 1 shows a plot of this function for a simulated situation with two factors. Given the $L_2-$norm restriction imposed on the coefficients $(\alpha_1, \alpha_2)$, all the feasible solutions in a two-dimensional setting lie on the unit circumference represented by $\alpha_1^2 + \alpha_2^2 = 1$, so each solution can be represented in spherical coordinates (a.k.a., polar coordinates in $\mathbb{R}^2$), by using only one variable: $\theta$.

The problem of multiple optima had not been faced by previous research efforts in the area (Martínez et al., 2016; Chudt & Nieto, 2018), since their approach was to pick among a finite set of alternatives (the set of estimated factors) a coincident index, instead of exploring over a set with uncountably many feasible points in $\mathbb{R}^r$.

To overcome this barrier, another element had to be incorporated into the objective function for the optimization problem. Based on the fact that Peña & Poncela (2006) and Martínez et al. (2016) handle with non-stationary time series

in their methodology, a good candidate to be part of the objective function is the cross-correlation at lag 0 between $\{\Delta\hat{c}_t\}$ and $\{\Delta i_t\}$: $cor\left(\Delta\hat{c}_t, \Delta i_t\right)$, assuming that the bivariate process $\left\{\left(\Delta\hat{C}_t, \Delta I_t\right)\right\}$ is stationary.

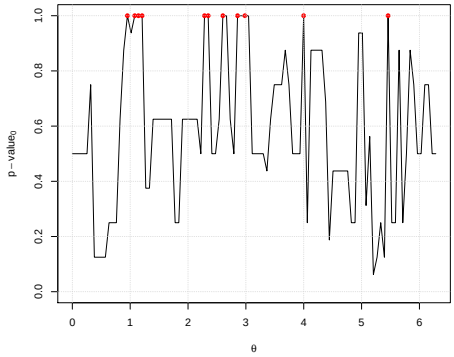



FIGURE 1: 0-coincidence $p$-value function for a two-dimension (2D) factor simulated / scenario.

The cross-correlation function allows to consider that, besides sharing the turning points with the proxy of the economy, the first difference of the index (analogous to the first derivative in a continuous approach) must share similar contemporary behavior with the first difference of the proxy for the economy. Chudt & Nieto (2018) use the cross-correlation at lag 0 as a criterion for selection of macroeconomic variables as indicators.

Because of the way the cross-correlation function is estimated, this only makes sense if both the $\{\Delta\hat{c}_t\}$ and the $\{\Delta i_t\}$ series come from stationary processes, i.e., if the original series $\{\hat{c}_t\}$ and $\{i_t\}$ are at most $I\left(1\right)$. In case the original series are integrated in a higher order, the cross-correlation function would have to be computed after applying as many difference operators as necessary to make the series stationary.

Additionally, this new element of the objective function lies between -1 and 1, which is relatively similar to the range in which the $p$-value for 0-coincidence varies (from 0 to 1), so there is no risk that one of the components of the objective reaches abnormally high values in magnitude to overshadow the other component. The two components, $p\left(\cdot,\cdot\right)$ and $cor\left(\cdot,\cdot\right)$, are included in the objective as the terms of a simple sum and the optimal solution will, then, exhibit appealing attributes with respect to both objectives.

The optimization problem in Equation (4) can be reformulated as

$$\begin{aligned} \max_{(\theta,\phi_1,\phi_2,\ldots,\phi_{r-2})} z &= \left[p\left(\Delta\hat{c}_t, \Delta i_t\right) + cor\left(\Delta\hat{c}_t, \Delta i_t\right)\right], \\ \text{s.t. } i_t &= \sum_{j=1}^{r} \alpha_j\left(\theta, \phi_1, \phi_2, \ldots, \phi_{r-2}\right) \cdot \hat{f}_{jt}, \\ & 0 \leq \phi_l \leq \pi, l \in \{1, 2, \ldots, r-2\}; \\ & 0 \leq \theta \leq 2\pi. \end{aligned} \tag{7}$$

This is a constrained optimization problem with an apparent non-convex objective function (as evidenced in multiple simulations) and a convex feasible region.

### 3.3. Use of a Genetic Algorithm to Reach the Optimal Value

The intractability of an analytical expression for the objective function reduces significantly the algorithms that can be implemented to find its optimal value because the vast majority of the methods are based on the possibility of computing at least first-order derivatives or sub-gradients of the objective function to define a good search direction, starting at an initial feasible point. The type of algorithms that exploit only the values for the objective function are denominated derivative-free algorithms (Rios & Sahinidis, 2013).

Both the Nelder-Mead and a genetic algorithm were tested with simulated results, yet, the genetic algorithm outperformed the Nelder-Mead algorithm in attaining a better feasible point. For that reason, the results presented in Section 6 were obtained using a genetic algorithm, by means of the package `GA` in `R` (Scrucca et al., 2013).

The parameters considered were: 10 generations, 100 individuals in each generation, 0.8 as the probability for crossover, and 0.1 as the probability for mutation.

## 4. Statistical Inference Based on the Linear-Combination Coefficient Estimators

The estimators of the unknown coefficients $\alpha_j \in \mathbb{R}, j = 1, 2, \ldots, r;$ are such that

$$\hat{\alpha}_j := \alpha_j \left( \hat{\theta}, \hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_{r-2} \right), \tag{8}$$

where

$$\left( \hat{\theta}, \hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_{r-2} \right) = \underset{(\theta, \phi_1, \phi_2, \ldots, \phi_{r-2})}{\arg\max} \; z = \left[ p \left( \Delta \hat{C}_t, \Delta I_t \right) + cor \left( \Delta \hat{C}_t, \Delta I_t \right) \right], \tag{9}$$

subject to the constraints presented in Equation (7).

Be aware of the changes from lower-case letters to upper-case letters to emphasize that the estimators are random variables because they depend on the underlying stochastic processes that generate the series included in the analysis. When particular realizations of these stochastic processes are considered, the arguments of the optimization problem are realizations of the random variables defined in Equation (9).

This procedure to obtain estimators via a non-conventional optimization technique is quite rare and it is very little what can be found in the literature about similar approaches.

It is known that these point estimates come from an optimal procedure that attempts to maximize the level of $0-$coincidence between a potential index and

a proxy for the state of the economy, which makes this estimation procedure very appealing for practitioners dealing with situations in which the DCFs and the coincident index framework applies. Nonetheless, it is important to consider mechanisms to construct confidence regions or test hypotheses regarding the coefficients of the linear combinations, since this is a way to confirm if the linear combination approach offers a useful generalization of previous approaches. For instance, assume that some estimates for the linear-combination coefficients, $\vec{\alpha}$, have been obtained following the methodology presented in this document, and it is of interest for the researcher to identify if the data support the hypothesis that only one of the DCFs is the most suitable coincident index that can be defined. This scenario can be translated into the following hypothesis system for some $j$, $j = 1, 2, \ldots, r$

$$\begin{cases} H_0 : (\alpha_1, \ldots, \alpha_j, \ldots, \alpha_r)^T = (0, \ldots, 1, \ldots, 0)^T \\ \text{v.s.} \\ H_1 : (\alpha_1, \ldots, \alpha_j, \ldots, \alpha_r)^T \neq (0, \ldots, 1, \ldots, 0)^T. \end{cases} \tag{10}$$

The null hypothesis in system (10) supports the idea that only the estimated $j$-th factor, for a particular $j$, is a coincident index for the state of the economy, being consistent with previous developments. On the other hand, if the null hypothesis is rejected for all $j$, $j \in \{1, 2, \ldots, r\}$; there is evidence to conclude that the coincident index cannot be composed of only one factor, meaning that a nontrivial linear combination can perform substantially better.

A simulation-based methodology is a sensible alternative to tackle the inferential problem posed by the hypothesis system in Equation (10) due to the inherent complexity of any theoretical approach. The methodology presented here can be easily extended to more general hypothesis systems.

Following the general representation of the DCF model, and assuming that for a given vector of macroeconomic time series of length $S$, $\{y_t\}_{t=1}^{S}$, the DCFs, $\left\{\hat{f}_t\right\}_{t=1}^{S}$, the matrices $\hat{\Sigma}_a$, $\hat{\Sigma}_e$ and $\hat{P}$ have been estimated; and also that the corresponding evolution structure for the dynamic factors, i.e., their *VARMA* structure, has been identified; it is possible to simulate new instances from that framework and generate different realizations of the linear-combination coefficients, by following the steps in Algorithm 1.

Step 1 in Algorithm 1 generates a simulated realization, $\{\tilde{y}_t\}_{t=1}^{S\ (n)}$, of the macroeconomic time series following the procedure in Algorithm 2. It is important to notice that each of the simulated vector time series $\{\tilde{y}_t\}_{t=1}^{S\ (n)}, n \in \{1, 2, \ldots, N\}$ for some simulation sample size $N$, is generated using the same parameters that were estimated from the real data at hand as input (see Algorithm 2). In case the normality assumption for the multivariate white noise processes (denoted as *MVN*: Multivariate Normal) does not hold, the step 1 in Algorithm 2 can be replaced for a nonparametric re-sampling routine based on the residuals obtained during the parameter estimation procedure.

**Algorithm 1:** Re-sampling routine for inference

**Result:** Sequence of coefficients $\left\{\hat{\vec{\alpha}}\right\}_{n=1}^{N}$ and objective values $\{\hat{z^*}\}_{n=1}^{N}$ for $N$ simulated instances

**Input:** *VARMA* parameters ($p$, $q$, $\hat{\Phi}(\cdot)$, $\hat{\Theta}(\cdot)$) and initial conditions for $\left\{\hat{f}_t\right\}$; $\{\hat{c}_t\}_{t=1}^{S}$, $\hat{\Sigma}_a$, $\hat{\Sigma}_e$, $\hat{P}$, $S$(sample size of original data), $N$ (simulation sample size);

$n := 1$;

**while** $n \leq N$ **do**

1. Generate a simulated instance of the macroeconomic variables $\{\tilde{y}_t\}_{t=1}^{S\ (n)}$ based on **Algorithm 2**;
2. Estimate the number, nature, dynamics and values of the DCFs $\left\{\hat{f}_t^{\star}\right\}_{t=1}^{S\ (n)}$; based on $\{\tilde{y}_t\}_{t=1}^{S\ (n)}$ (Peña & Poncela, 2006);
3. Estimate the linear-combination coefficients $\left\{\hat{\vec{\alpha}}\right\}^{(n)}$ by solving (7) using $\{\hat{c}_t\}_{t=1}^{S}$ as proxy;
4. Compute and store the value of the objective function $\hat{z^*}^{(n)} := (\hat{p} + \hat{cor})^{(n)}$;
5. $n \leftarrow n + 1$;

**end**

Once a new multivariate time series, $\{\tilde{y}_t\}_{t=1}^{S\ (n)}$, has been simulated; it is necessary to repeat the procedure of estimation of the DCFs and then, identify the optimal combination of factors. The term $(\hat{p} + \hat{cor})^{(n)}$ denotes the estimate of the objective function in each draw of the $N$ simulations.

It would seem intuitive to use the entire sequence of realizations for the vector of coefficients, $\left\{\hat{\vec{\alpha}}^{(n)}\right\}_{n=1}^{N}$, to draw inferences about the population coefficients. Nevertheless, this approach would show, with no doubt, misleading conclusions in terms of inference about the parameters. The reason behind this lies in the way the simulation is conducted and the properties of the DCF model.

There is an analogy between DCF models and the principal component analysis in multivariate statistics. If $v$ is an orthonormal eigenvector of the matrix $\Sigma$ corresponding to the eigenvalue $\lambda$, the vector $-v$ will also be an orthonormal eigenvector of the matrix with the same eigenvalue. In the context of DCF models, the factors are analogous to the principal components of a multivariate random vector. In their estimation procedure, sometimes, it might result an estimation similar to $f_{jt}$ and in other cases similar to $-f_{jt}$ for some $j$, because both represent the same component's variability. Additionally, in the case of the classical principal component analysis, the sorted eigenvalues can help to identify the natural order of the principal components. In the case of the DCF models, there is not a unified criterion to sort the factors and pretty often, the factor that came first in a previous simulation instance can appear in a different position in the following instance.

**Algorithm 2:** Iterative simulation of macroeconomic variables

**Result:** New realization $\{\tilde{y}_t\}_{t=1}^{S}$ of same size as original data
**Input:** *VARMA* parameters $(p,\ q,\ \hat{\Phi}(\cdot),\ \hat{\Theta}(\cdot))$ and initial conditions for $\left\{\hat{f}_t\right\}$; $\hat{\Sigma}_a$, $\hat{\Sigma}_e$, $\hat{P}$, $S$(sample size of original data);
$s := 1$;
**while** $s \leq S$ **do**
1. Simulate the vector of errors: $\tilde{e}_s \leftarrow MVN\left(0, \hat{\Sigma}_e\right)$ and $\tilde{a}_s \leftarrow MVN\left(0, \hat{\Sigma}_a\right)$;
2. Compute the factors: $\tilde{f}_s$ such that $\hat{\Phi}(B)\,\tilde{f}_s = \hat{\Theta}(B)\,\tilde{a}_s$;
3. Compute the realization of the macroeconomic variables: $\tilde{y}_s \leftarrow \hat{P}\tilde{f}_s + \tilde{e}_s$;
4. $s \leftarrow s + 1$;

**end**

For these reasons, inference about the linear-combination coefficients cannot be drawn simply by considering the sequence $\left\{\hat{\tilde{\alpha}}^{(n)}\right\}_{n=1}^{N}$ as the realization of an i.i.d. sample of the multivariate estimator of the coefficients. This motivates a different use of the simulation routine.

Regardless of the ordering or the inverted sign of some factors, the objective function $\hat{z^*}^{(n)} = (\hat{p} + c\hat{o}r)^{(n)}$ gathers all the potential ability of that realization of the multivariate econometric time series to create an index that is 0-coincident with the proxy for the state of the economy. In that sense, the sequence $\left\{\hat{z^*}^{(n)}\right\}_{n=1}^{N}$ can be considered as a realization of a random sample that can be used to draw inferences about the population parameter $z^*$ (which is a function of the underlying stochastic processes).

Bickel & Doksum (2015) express that, given a confidence region $CI_{1-\delta}(\xi)$ for a generic parameter $\xi$ with a given level of confidence $1-\delta, \delta \in (0,1)$; a confidence region with the same level of confidence for a function of the parameter $q(\xi)$, can be defined as

$$CI_{1-\delta}(q(\xi)) := \{q(x) |\, x \in CI_{1-\delta}(\xi)\}.$$

There are no restrictions for the type of function $q(\cdot)$ that can be considered.

An analogous result was applied to the problem to be able to find a confidence region for the linear-combination coefficients based on a confidence interval for the objective function. A unilateral confidence interval for the objective function was considered for two reasons. Firstly, the objective function is bounded (it has an upper bound of 2). Secondly, there is no reason to exclude values in the right tail (close to the upper bound) of the sample if the main purpose of the optimization is to maximize the objective function.

Considering that a $100 \cdot (1-\delta)\%$-confidence interval for the objective function $z^*$ has the general form $[L, 2]$, $L \geq 0$, $\delta \in (0, 1)$; and considering that there exists a function $g(\cdot)$ such that $g(\vec{\alpha}) = z^*$, a confidence region of level $1-\delta$ for the vector of coefficients $\vec{\alpha}$ corresponds to the set

$$\left\{\vec{\alpha} \in \mathbb{R}^r : g(\vec{\alpha}) \in [L, 2] \text{ and } \|\vec{\alpha}\|_2^2 = 1\right\},$$

because the following equivalences hold in the common probability space $(\Omega, \Im, \mathbb{P})$ that allows all the random applications involved to be proper random variables

$$\begin{aligned} 1-\delta &= \mathbb{P}\{\omega \in \Omega : [L(\omega), 2] \ni z^*\} \\ &= \mathbb{P}\{\omega \in \Omega : [L(\omega), 2] \ni g(\vec{\alpha})\} \\ &= \mathbb{P}\left\{\omega \in \Omega : g^{-1}([L(\omega), 2]) \ni \vec{\alpha}\right\}, \end{aligned}$$

where $g^{-1}([L(\omega), 2])$ is the set of pre-images in $\mathbb{R}^r$ that under $g(\cdot)$ fall within $[L, 2]$.

In the application of this result, it has to be noted that $g(\vec{\alpha}) = p(\Delta\hat{c}_t, \Delta i_t(\vec{\alpha})) + cor(\Delta\hat{c}_t, \Delta i_t(\vec{\alpha}))$, but there is no analytical expression for the function $g^{-1}(\cdot)$; so, once a confidence interval for the objective value of the problem has been computed, a numerical approximation of the set of pre-images has to be performed. On the other hand, when used for hypothesis testing for a candidate null hypothesis of the form $H_0 : \vec{\alpha} = \vec{\alpha}_0$, it suffices to verify whether or not $g(\vec{\alpha}_0) \in [L, 2]$ to decide if that null hypothesis can be rejected or not at a level of significance of $100\hat{\delta}\%$.

# 5. Artificially-Generated Scenarios for Validation

To check the usefulness of the methodology proposed in this work, some randomly generated scenarios are presented to highlight its advantages in effectively tackling the challenges that the nature of the problem poses. The procedure to generate the scenarios is briefly described in the following Subsection.

## 5.1. Generation of a Base Scenario

The generation framework of a base condition was motivated by Stock et al.'s approach (Stock & Watson, 1992), in the sense that the state of the economy is the main common feature among the macroeconomic variables. The state of the economy, $\{C_t\}$, was assumed to follow a general $ARIMA$ model, i.e., $\{C_t\} \sim ARIMA(p, d, q)$.

According to Stock et al.'s model, $\{C_t\}$ is the input to generate the vector of macroeconomic variables (there is no mediation of multiple DCFs). However, for the sake of the analysis and to be able to fully implement the methodology previously described, it was assumed that the state of the economy generates a vector of $r$ common factors via the expression

$$f_t = P_f C_t + u_t,$$

where $u_t \sim MVN\left(0, I_{r\times r}\right)$, being $I_{r\times r}$ the identity matrix of rank $r$. On the other hand, the vector of $m$ macroeconomic variables was generated via the expression

$$Y_t = Pf_t + e_t,$$

with $e_t \sim MVN\left(0, \Sigma_e\right)$ and $\Sigma_e = \text{diag}\left\{{}^{i^2}\!/_{h\left(m\right)} : 1 \leq i \leq m\right\}$.

The structure presented for $\Sigma_e$ attempts to introduce some heterogeneity with a scaling factor, $h(m)$, to avoid excessively large variances. In the simulations presented $h(m) = m - 1$.

As it was stated before, the intermediate step of calculating the factors was included to be able to compare the generated factors with the estimations obtained with the Kalman Filter.

Another important observation is that, in order to be able to construct the scenarios, a pre-specified dimension, $r$, for the vector of factors is used as given. However, this does not necessarily imply that the variables exhibit as many common trends as that dimension size because the matrices $P$ and $P_f$ are determined in a random fashion, implying that the number of effective DCFs present in each scenario can vary from 1 to $r$.

Once a full set of realizations for the state of the economy, the vector of DCFs and the vector of macroeconomic variables has been obtained, the methodology presented in Sections 3 and 4 can be applied. In that regard, it was assumed that only the realizations of the vector of macroeconomic variables and the state of the economy were available. The factors generated were only considered to compare to the estimated factors obtained from the macroeconomic variables.

## 5.2. Estimation of the DCFs and Inference

Firstly, it would be necessary to estimate the number, nature and values of the DCFs based on the macroeconomic variables available (Peña & Poncela, 2006). For the estimation of the factors, the `MARSS` package in `R` was used (Holmes et al., 2012). Since the objective of this work is not to deal with the issues in the identification of the number of factors, it was assumed that the number of factors was equal to $r$ (the parameter used in the generation, Subsection 5.1).

The identification and estimation of the dynamics according to which the factors evolve in time, i.e., their *VARIMA* structure; is another aspect that might be determined based on the data. However, for the same reason as before, the factors were assumed to always follow a multivariate random walk dynamics, $f_t = f_{t-1} + a_t$, where $a_t \sim MVN\left(0, \Sigma_a\right)$ and $\Sigma_a$ is a diagonal matrix. This assumption is common in the state-space models because, although simple, it allows to fairly capture general evolution dynamics (Hamilton, 1994),(Pivetta & Reis, 2007).

Once the factors for the model have been estimated, it is possible to estimate coefficients of the best coincident index in terms of its 0-coincidence and its cross-correlation at lag 0 with the realization of the proxy for the state of the economy.

For illustrative purposes, the proxy available was considered to be the realization of the state of the economy itself, although this is not possible in real applications for obvious reasons.

Finally, by following the procedures presented in Algorithm 1, it was possible to conduct inference about the linear-combination coefficients, i.e., to create a confidence region with a certain level of confidence and to assess specific systems of hypothesis.

# 6. Results

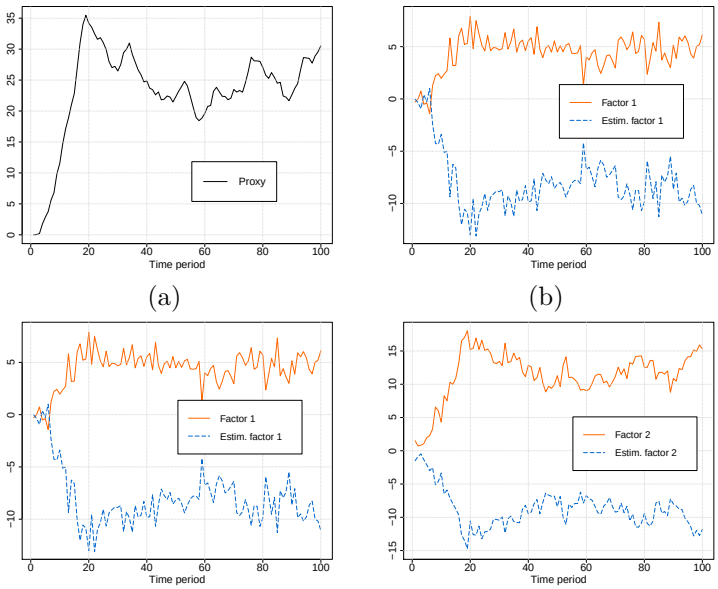
This section presents the results of some artificially-generated scenarios as well as an application to real data to show how the methodology already described can be applied. All the analyses carried out in this work were developed using the software `R` (v4.1.2, R Core Team, 2021). Two types of artificial instances were generated: a 2D-type (two-factor instance) and a 3D-type (three-factor instance). A sample size of 100 temporal observations was used for all the generated scenarios, which is a reasonable amount compared to a situation in which variables are measured monthly over a time window of 8 years approx. Finally, the methodology was applied to the results previously obtained by Chudt & Nieto (2018).

## 6.1. Results for 2D

For this instance, it was assumed that the number of DCFs was 2, $r = 2$, and the number of variables was $m = 5$. Additionally, the state of the economy was assumed to follow an $ARIMA\,(1, 1, 0)$ model with autoregressive parameter $\phi = 0.7$. The matrices $P_f$ and $P$ were generated by a random mechanism considering the identiability requirements. For this particular scenario, the matrices were

$$P_f = \begin{bmatrix} 0.18 \\ 0.49 \end{bmatrix}, P = \begin{bmatrix} 0.49 & -0.98 & -6.03 & -0.91 & 3.53 \\ 0 & -2.15 & -1.39 & 2.87 & -5.22 \end{bmatrix}^T.$$

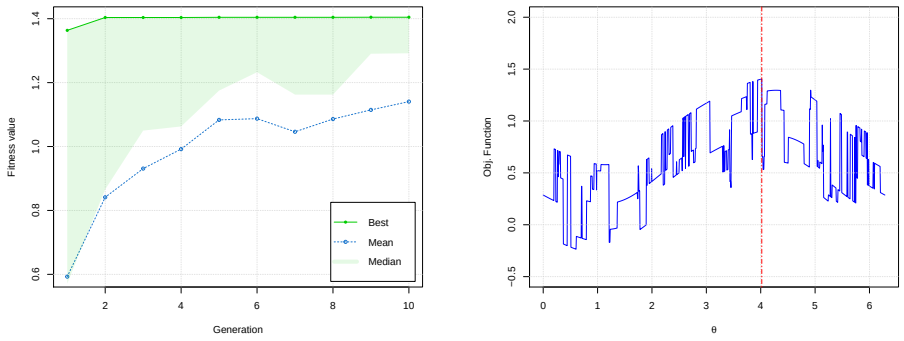
Once the state of the economy, the factors and the macroeconomic variables were generated, the realization of the macroeconomic variables was used to estimate the 2 factors using the Kalman Filter. The results for the generated and the estimated factors are presented in Figure 2. It is important to notice that, as it was pointed out before, sometimes the estimates of the factors can have a different sign with respect of the original factors, highlighting the importance of including linear combinations (in some cases with negative coefficients) to build a coincident index.




(a) (b)

FIGURE 2: (a) Plot of the proxy generated, (b-c) Comparison of the factors generated versus the estimated ones

Figure 3(a) shows the overall progress through the generations of the optimization subroutine based on a genetic algorithm. Even if there is not a given criterion of convergence for the genetic algorithm procedure (in terms of closeness to the optimal solution), the fact that the best solution found does not significantly change in the last generations, and the best value, the median and the mean values along the generations seem to approach to each other are good signs of the quality of the solution in terms of optimality.




(a) (b)

FIGURE 3: (a) Progress monitoring of the genetic algorithm subroutine (2D). For each generation, the best, the median and the mean value of the individuals of that generation are presented, (b) Objective function in spherical coordinates (optimal solution highlighted).

The results show that the optimal value for this case is $\hat{z}^* = 1.405$, corresponding to a 0-coincidence $p$-value of 1 and a cross-correlation of 0.405. This optimal value is attained at $\hat{\theta} = 4.014$ in spherical coordinates, and $\hat{\alpha_1} = -0.6387, \hat{\alpha_2} = -0.769$ in Cartesian coordinates. According to this result, the coincident index is close to an average of the negatives of the two estimated factors. Figure 3b corresponds to a plot of the objective function in spherical coordinates, highlighting the value obtained from the optimization routine, while Figure 4 shows the comparison between the proxy and the coincident index created as a linear combination of the estimated factors.

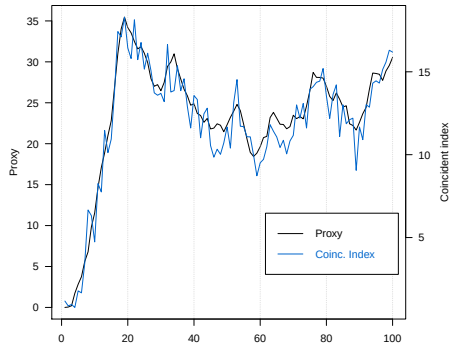



FIGURE 4: Comparison between the realization of the state of the economy (proxy) and the coincident index (2D).

There is a different scale for the proxy and the coincident index because the idea behind the construction of a coincident index is to try to replicate the fluctuations that the state of the economy experiences rather than replicating the set of values it takes. It can be seen how the coincident index fairly captures the dynamics of the state of the economy. Based on the resampling technique proposed, a 95% unilateral confidence interval for the optimal value of the objective function is $[1.001, 2]$. The histogram for 500 realizations of the objective function is presented in Figure 5.

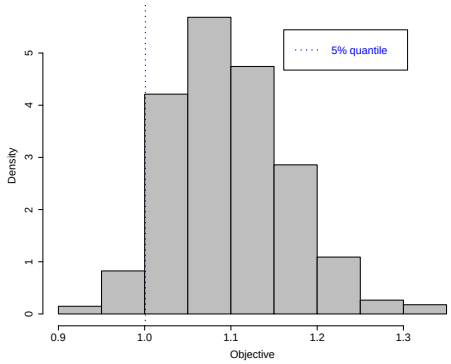
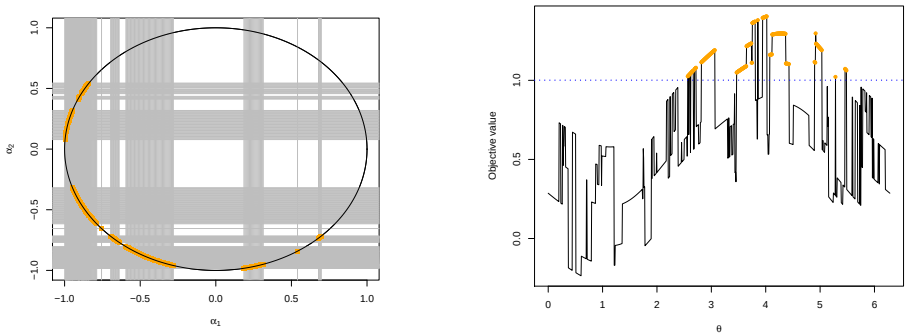



FIGURE 5: Histogram of the realizations for the optimal value of the objective function.

Once the confidence interval for the objective's optimum has been estimated, it is possible to create a 95% confidence region for the coefficients of the linear combination in both Cartesian and spherical coordinates, as shown in Figure 6. According to the results of the confidence regions, it can be concluded that no single factor can actually be a coincident index.




(a) (b)

FIGURE 6: (a) Confidence region in Cartesian coordinates (2D). The highlighted area over the unit circumference represents the 95% confidence region while the straight lines represent projections onto the axes. (b) Confidence region in spherical coordinates (2D). The vertical threshold corresponds to the lower limit of the confidence interval while the horizontal coordinate of the highlighted points represents the confidence region for the spherical argument.

## 6.2. Results in 3D

In this scenario, it was assumed that the number of DCFs was 3, $r = 3$, and the number of variables was $m = 7$. Additionally, the state of the economy was assumed to follow an $ARIMA(1, 1, 0)$ model with $\phi = 0.7$. For this scenario, the matrices $P_f$ and $P$ were

$$P_f = \begin{bmatrix} -0.92 \\ 0.24 \\ -0.79 \end{bmatrix}, P = \begin{bmatrix} 7.7 & -1.26 & 6.18 & -1.94 & 7.94 & -4.37 & 1.72 \\ 0 & -6.19 & 2.5 & -6.24 & 0.73 & -4.41 & 8.29 \\ 0 & 0 & -3.89 & 2.35 & -4.02 & -0.91 & 2.76 \end{bmatrix}^T.$$

The results for the generated and the estimated factors are presented in Figure 7. This scenario shows a characteristic worth revising: the order in which the factors are estimated and presented by the Kalman Filter does not necessarily correspond to the order in which they were originally generated. For instance, factor 1 and factor 3 seem to have been misplaced during the estimation. It is important to remind that in a real situation, it is not possible to know in which order the factors are estimated.

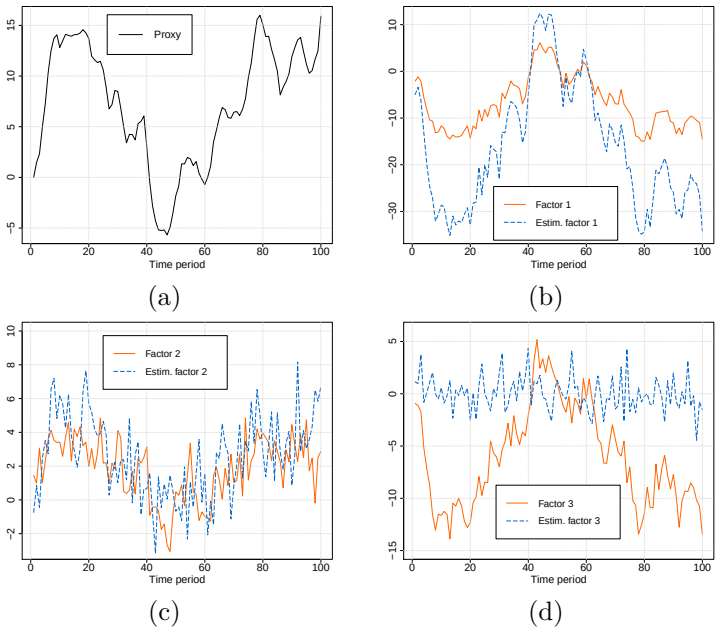



FIGURE 7: (a) Plot of the proxy generated, (b-d) Comparison of the factors generated versus the estimated ones

After running the genetic algorithm subroutine, the optimal value for this case is $\hat{z^*} = 1.711$, corresponding to a 0-coincidence $p$-value of 1 and a cross-correlation of 0.711. This optimal value is attained at $\hat{\theta} = 1.495, \hat{\varphi} = 2.781$ in spherical coordinates, and $\hat{\alpha_1} = -0.933, \hat{\alpha_2} = 0.352, \hat{\alpha_3} = 0.075$ in Cartesian coordinates. Figure 8 shows the comparison between the realization of the state of the economy and the coincident index created as a linear combination of the estimated factors.

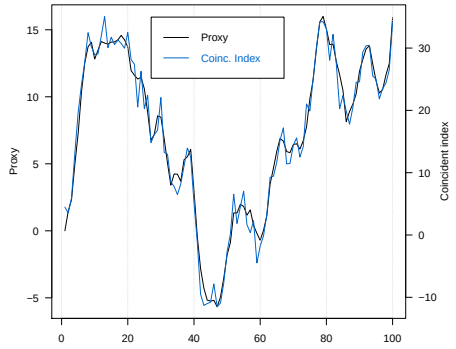



FIGURE 8: Comparison between the realization of the state of the economy (proxy) and the coincident index (3D).

Based on the resampling technique proposed, a 95% unilateral confidence interval for the optimal value of the objective function is $[1.097, 2]$. Figure 9 shows

the confidence region in spherical coordinates. According to the results of the confidence region, it can be concluded that factor 2 could be a coincident index for the state of the economy since $(\theta, \varphi) = \left(\frac{\pi}{2}, \frac{\pi}{2}\right)$ belongs to the 95% confidence region, as well as the negative of factor 1 represented by $(\theta, \varphi) = \left(\pi, \frac{\pi}{2}\right)$.

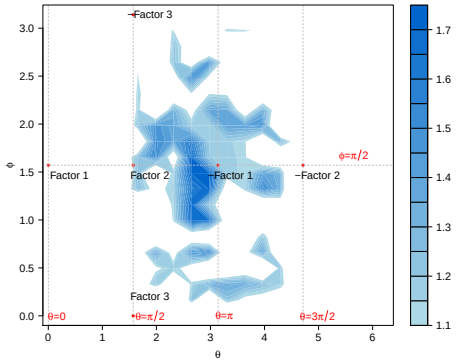



FIGURE 9: Confidence region in spherical coordinates (3D).

## 6.3. Aggregate Performance on Multiple Generated Instances

The procedure presented in Subsections 6.1 and 6.2 was replicated 100 times starting from random and independent base case scenarios for both the 2D and the 3D scenarios.

For each one of the replications, the following information was recorded: (1) whether the hypothesis system for all possible trivial combinations was rejected or not, i.e., if a non-trivial linear combination was the best attempt to model the behavior of the state of the economy; (2) the gap in terms of the objective function value for the non-trivial linear combination estimated and the highest value of the objective among the single factors

$$gap := z^* - \max_{l \in \{1,2,\ldots,r\}} \left[ p\left(\Delta\hat{c}_t, \Delta\hat{f}_{lt}\right) + cor\left(\Delta\hat{c}_t, \Delta\hat{f}_{lt}\right) \right], \tag{11}$$

and (3) the relative improvement of the gap (expressed as a percentage) in terms of the highest value of the objective among the single factors

$$RI := \frac{gap}{\max_{l \in \{1,2,\ldots,r\}} \left[ p\left(\Delta\hat{c}_t, \Delta\hat{f}_{lt}\right) + cor\left(\Delta\hat{c}_t, \Delta\hat{f}_{lt}\right) \right]}. \tag{12}$$

These quantities were then summarized for the 100 replications (in terms of mean values) and are presented separately for the cases in which a non-trivial linear combination was the best alternative versus the cases in which that does not hold true based on the hypothesis tests. The results can be seen in Table 1.

TABLE 1: Aggregate performance summary for the replications

| *Scenario* | *Number of replications* | *% non-trivial* | $\bar{gap}$ | $\bar{RI}$ | *%trivial* | $\bar{gap}$ | $\bar{RI}$ |
|---|---|---|---|---|---|---|---|
| *2D* | 100 | 44% | 0.345 | 33.85% | 56% | 0.182 | 17.14% |
| *3D* | 100 | 61% | 0.522 | 41.76% | 39% | 0.207 | 20.08% |

For instance, for the 2D scenario, 44% of the replications showed that a non-trivial linear combination of factors was the best coincident index to replicate the proxy. Among this 44% of the cases, the average improvement of the objective value ($p$-value for 0 coincidence plus cross-correlation at lag 0) of the non-trivial combination with respect to the best single-factor choice is 0.345. In relative terms, the average improvement is close to the 34%. For the remaining 56% of the cases, the gap and the relative improvement are obviously lower because the data support the idea that one of the single factors itself can play the role of the coincident index. In an analogous way, for the 3D scenario, the majority of the replications (61%) showed that non-trivial linear combinations performed better than single factors in terms of the objective value. The margins that account for the performance difference were higher too (average gap of 0.522 and average relative improvement of 42% approx.).

## 6.4. Application to the Colombian Economy

Chudt & Nieto (2018) computed a new coincident index for the Colombian economy based on the following six monthly macroeconomic variables: (1) Industrial Production Index (IP), (2) Average Electric Energy Consumption (EEC) in GWh/day, (3) Retailing Commerce Index Excluding Fuels and Vehicles (RC), (4) Total Production of Sugar Cane (SCP) in tons, (5) Cement Production (CP) in tons, and (6) Unemployment Rate (UR) as a percentage; during the period starting in January, 2000 until June, 2017. The six indicators are presented in Figure 10.

After deseasonalizing and pre-processing the series according to Martínez et al. (2016), they identified two DCFs for the macroeconomic series and used the Economic Tracking Index (ISE for its initials in Spanish: Índice de Seguimiento Económico) as a proxy for the state of the economy. The ISE is computed by the Official Statistics Bureau for Colombia (DANE for its acronym in Spanish) on a monthly basis since January, 2000.

Figure 11 shows the (deseasonalized) proxy for the state of the economy and the two estimated factors.

The optimization process to estimate the coefficients of a linear combination of the two factors led to the following results: the objective value was $z^* = 1.201$, corresponding to a 0-coincidence $p$-value of 0.875 and a correlation of 0.326. This optimal value is attained at $\hat{\theta} = 6.282$ in spherical coordinates, and $\hat{\alpha_1} = 0.999$, $\hat{\alpha_2} = -0.002$ in Cartesian coordinates. The results obtained suggest that the coincident index is primarily composed of factor 1. To test that, the 95%-confidence interval for the objective's optimum was calculated giving as result $[0.85, 2]$. With

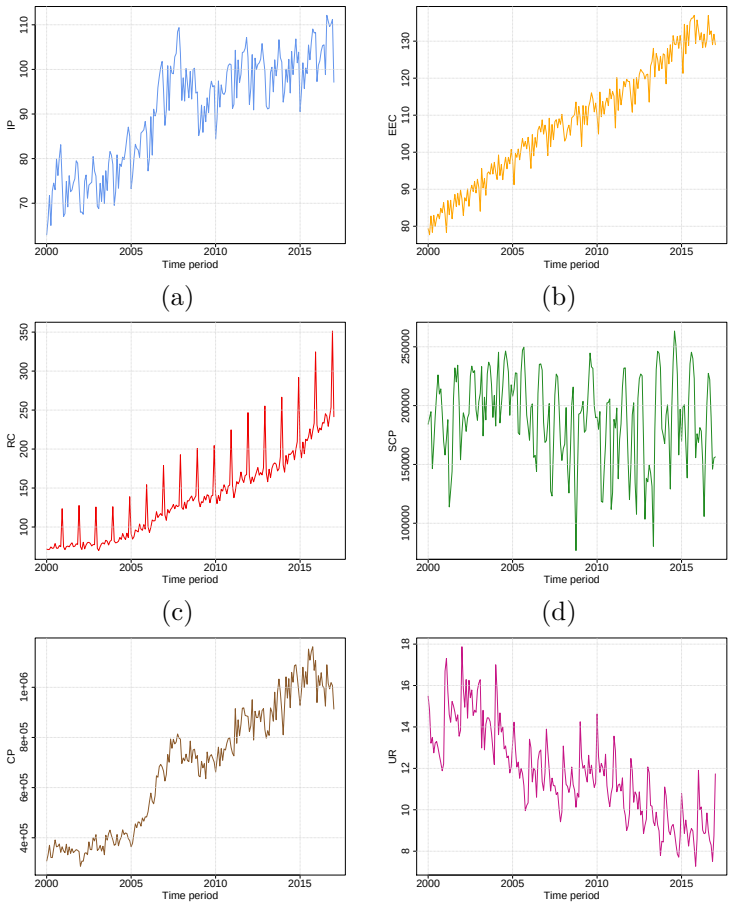



FIGURE 10: Macroeconomic indicators used to create the coincident index. (a) IP, (b) EEC, (c) RC, (d) SCP, (e) CP, (f) UR.

this result, the 95%-confidence region for the coefficients of the linear combination is presented in Figure 12. Based on this confidence region, it can be concluded that, with a significance level of 5%, factor 1 (with coefficients $(\alpha_1, \alpha_2) = (1, 0)$) can play the role of the coincident index.

The results for the application in the Colombian context are consistent to what had been previously obtained by Chudt & Nieto (2018) because both procedures have suggested that factor 1 can play the role of the coincident index for the Colombian economy. Nevertheless, the primary value that the new methodology offers is the possibility to expand the set of candidates to be coincident index and actually conclude, via a statistical test, that among all the possible linear combinations of factors, factor 1 can be considered as the most suitable choice for coincident index.

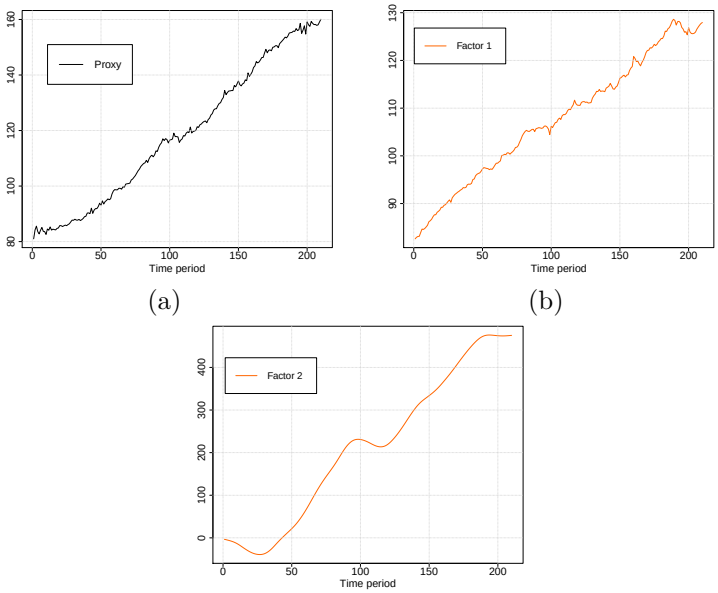



FIGURE 11: (a) Deseasonalized proxy (ISE), (b-c) Estimated factors.

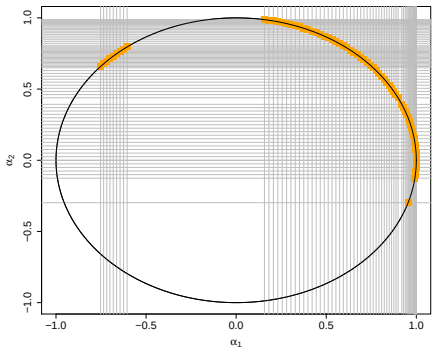



FIGURE 12: Confidence region in Cartesian coordinates for the Colombian case study.

# 7. Conclusions

This work proposes a novel methodology to design coincident indices as linear combinations of the dynamic common factors from a multivariate time series of macroeconomic variables with the advantage of directly and effectively handling potential nonstationarity in the series, unlike other approaches in the literature. The first and one of the key steps in the methodology proposed is to estimate the dynamic common factors from the macroeconomic time series, to capture the common trends and patterns in a lower-dimension object, which is particularly advantageous since the dimension of the vector of factors determines the com-

plexity of the optimization problem that needs to be solved in order to estimate the coefficients of the optimal linear combination of factors. Using directly the macroeconomic variables to construct a coincident index would cause the genetic algorithm to grow exponentially in complexity and consume an excessive amount of computation time.

The linear-combination coefficients are estimated based on the objective of maximizing the 0-coincidence between the candidate index and a proxy for the state of the economy. Doing this guarantees that the coincident index obtained provides the best match with the dynamics of the proxy in terms of its peaks and its valleys and such information is of utmost importance for economic planning since it characterizes periods of prosperity and recession. The coincident index is not designed to nowcast the values of the proxy per se, which justifies the use of the 0-coincidence $p$-value as the optimization objective, instead of using least squares or other loss functions to estimate the linear-combination coefficients. However, the inclusion of the contemporary cross-correlation of the candidate index and the proxy as part of the objective allows considering, to some extent, the similarity in the values of the two series. This work also provides a set of statistical inference procedures to compute, in addition, interval estimates and test hypotheses regarding the linear-combination coefficients.

Even if this new methodology is presented as an improvement of one of the phases in Martinez et al.'s methodology (Martínez et al., 2016), it can be easily adapted and linked to other methodologies to estimate DCFs in multivariate time series, such as Forni et al. (2005), Stock & Watson (2011), Lam et al. (2012), Bujosa et al. (2013); among others.

On the other hand, the simulated scenarios establish several facts about the importance of this work because, as it could be seen, (1) there are situations in which a single factor cannot explain the dynamics in the business cycle; and (2) the estimated factors usually come with opposite signs, requiring the use of negative multiplicative constants to actually mimic the behavior of the proxy. Finally, the methodology was applied to the Colombian context producing consistent results to what had been previously obtained, and offering stronger statistical evidence to support the conclusion of Chudt & Nieto (2018), by means of the analysis of the sampling distribution for the linear-combination estimators.

## Acknowledgements

The authors would like to thank Universidad Nacional de Colombia for the support through the "Beca Grado de Honor" scholarship, which allowed Dr. Arrieta-Prieto to pursue his master's degree in statistics. This publication is based on his master's thesis (Arrieta-Prieto, 2019).

[Received: June 2024 — Accepted: August 2024]

## References

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M. & Veronese, G. (2010), 'New Eurocoin: tracking economic growth in real time', *The Review of Economics and Statistics* **92**(4), 1024–1034.

Arrieta-Prieto, M. E. (2019), 'Selection of a linear combination of common factors as a coincident index for the Colombian economy'. https://repositorio.unal.edu.co/handle/unal/69244

Banerji, A. (1999), The lead profile and other non-parametric tools to evaluate survey series as leading indicators, *in* 'Use of Survey Data for Industry, Research and Economic Policy, selected papers presented at the 24th CIRET Conference, Wellington, New Zealand'.

Bazaraa, M. S., Sherali, H. D. & Shetty, C. M. (2013), *Nonlinear programming: theory and algorithms*, John Wiley & Sons.

Bickel, P. J. & Doksum, K. A. (2015), *Mathematical statistics: basic ideas and selected topics, volume I*, Vol. 117, CRC Press.

Blumenson, L. (1960), 'A derivation of n-dimensional spherical coordinates', *The American Mathematical Monthly* **67**(1), 63–66.

Bujosa, M., García-Ferrer, A. & Juan, A. (2013), 'Predicting recessions with factor linear dynamic harmonic regressions', *Journal of Forecasting* **32**(6), 481–499.

Burns, A. F. & Mitchell, W. C. (1946), Measuring business cycles, Technical report, National Bureau of Economic Research, Inc.

Chudt, N. & Nieto, F. (2018), 'Construcción de un índice coincidente para la actividad económica Colombiana', *Comunicaciones en Estadística* **11**, 107–128.

Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2005), 'The generalized dynamic factor model: one-sided estimation and forecasting', *Journal of the American Statistical Association* **100**(471), 830–840.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton: Princeton university press.

Holmes, E. E., Ward, E. J. & Wills, K. (2012), 'MARSS: Multivariate Autoregressive State-space Models for Analyzing Time-series Data', *R Journal* **4**(1).

Lam, C., Yao, Q. et al. (2012), 'Factor modeling for high-dimensional time series: inference for the number of factors', *The Annals of Statistics* **40**(2), 694–726.

Lütkepohl, H. (2005), *New introduction to multiple time series analysis*, Springer Science & Business Media.

Martínez, W., Nieto, F. H. & Poncela, P. (2016), 'Choosing a dynamic common factor as a coincident index', *Statistics & Probability Letters* **109**, 89–98.

Nieto, F. H., Pena, D. & Saboyá, D. (2016), 'Common seasonality in multivariate time series', *Statistica Sinica* pp. 1389–1410.

Peña, D. & Poncela, P. (2006), 'Nonstationary dynamic factor analysis', *Journal of Statistical Planning and Inference* **136**(4), 1237–1257.

Peña, D. & Tsay, R. S. (2021), *Statistical learning for big dependent data*, John Wiley & Sons.

Pivetta, F. & Reis, R. (2007), 'The persistence of inflation in the United States', *Journal of Economic Dynamics and Control* **31**(4), 1326–1358.

Rios, L. M. & Sahinidis, N. V. (2013), 'Derivative-free optimization: a review of algorithms and comparison of software implementations', *Journal of Global Optimization* **56**(3), 1247–1293.

Scrucca, L. et al. (2013), 'GA: a package for genetic algorithms in R', *Journal of Statistical Software* **53**(4), 1–37.

Stock, J. H. & Watson, M. W. (1992), 'A probability model of the coincident economic indicators', *Leading Economic Indicators: New Approaches and Forecasting Records* p. 63.

Stock, J. H. & Watson, M. W. (2011), Dynamic factor models, *in* 'The Oxford Handbook of Economic Forecasting'.

v4.1.2, R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Wei, William WS (2006), *Time series analysis*, Pearson Addison Wesley.