

*Selection of a Linear Combination of Common Factors as  
a Coincident Index for the Colombian Economy*

MARIO ENRIQUE ARRIETA PRIETO  
STATISTICIAN, INDUSTRIAL ENGINEER



UNIVERSIDAD NACIONAL DE COLOMBIA  
FACULTAD DE CIENCIAS  
DEPARTAMENTO DE ESTADÍSTICA  
BOGOTÁ , D.C.  
NOVEMBER, 2018

*Selection of a Linear Combination of Common Factors as  
a Coincident Index for the Colombian Economy*

MARIO ENRIQUE ARRIETA PRIETO  
STATISTICIAN, INDUSTRIAL ENGINEER

A DISSERTATION SUBMITTED FOR THE DEGREE OF  
MASTER OF SCIENCE, STATISTICS

ADVISOR  
FABIO HUMBERTO NIETO SÁNCHEZ, PH.D.  
PH.D. IN STATISTICS

RESEARCH LINE  
TIME SERIES ANALYSIS

RESEARCH GROUP  
TIME SERIES



UNIVERSIDAD NACIONAL DE COLOMBIA  
FACULTAD DE CIENCIAS  
DEPARTAMENTO DE ESTADÍSTICA  
BOGOTÁ , D.C.  
NOVEMBER, 2018

**Title in English**

Selection of a Linear Combination of Common Factors as a Coincident Index for the Colombian Economy

**Título en español**

Selección de una Combinación Lineal de Factores Comunes como un Índice Coincidente para la Economía Colombiana

**Abstract:** The main goal of this work is to propose a general methodology to create a coincident index based on linear combinations of common factors, and to test it on scenarios from simulations and on a case study for the Colombian economy. A whole methodological approach to produce point estimates, confidence regions, and to test hypotheses is presented. Besides, the results for the scenarios show how promising this new proposal is with respect to previous achievements in the scientific community.

**Resumen:** El principal objetivo de este trabajo es proponer una metodología general para crear un índice coincidente basado en combinaciones lineales de factores comunes, y aplicarla en escenarios simulados y un caso de estudio para la economía colombiana. Se presenta un enfoque metodológico completo para producir estimaciones puntuales, regiones de confianza y para juzgar hipótesis. Adicionalmente, los resultados de los escenarios muestran cuán prometedora es esta nueva propuesta con respecto a logros anteriores en la comunidad científica.

**Keywords:** Coincident index, Multivariate time series, Common dynamic factors, Coincident profile, Derivative-free optimization, Genetic algorithms.

**Palabras clave:** Índice coincidente, Series de tiempo multivariadas, Factores comunes dinámicos, Perfil coincidente, Optimización libre de derivadas, Algoritmos genéticos.

# Acceptation Note

Thesis Work

Approved

“Meritorious mention”

---

Jury

Ph.D. Sergio Alejandro Calderón Villanueva

---

Jury

M.Sc. Luis Fernando Melo Velandia

---

Advisor

Ph.D. Fabio Humberto Nieto Sánchez

Bogotá , D.C., November 27th, 2018

---

---

## Dedicated to

---

---

*I would like to dedicate this work to my mother, who was my greatest supporter. Now that she is gone, her memory and example are still my strongest source of motivation and inspiration to keep going.*

---

---

## Acknowledgements

---

---

First, I would like to thank my family for being such a strong support during this process. Especially my mom, who started this journey with me and will accompany me forever in my heart; my sister, who has always given me a constant boost of bravery and support during the last years; and my friends, who have played a key role in my life as the family I chose.

I am also in debt with my advisor, Prof. Fabio H. Nieto, for all his guidance and infinite patience throughout all this process.

I am very grateful to have had the chance to interact with excellent professors during my master's program in the Statistics Department and to have learnt such a solid foundation in statistics. Additionally, I would like to thank the university in general for the *honorary degree scholarship* grant, which allowed me to focus on my formation and my research efforts.

---

---

# Contents

---

---

|   |            |
|---|------------|
| <b>Contents</b>   | <b>I</b>   |
| <b>List of Tables</b>   | <b>II</b>  |
| <b>List of Figures</b>  | <b>III</b> |
| <b>Introduction</b>   | <b>V</b>   |
| <b>1. Literature Review</b>   | <b>1</b>   |
| <b>2. Methodology</b>   | <b>7</b>   |
| 2.1 Formulation of the Problem . . . . .  | 7          |
| 2.2 Use of Spherical Coordinates . . . . .  | 9          |
| 2.3 Inconveniences with the 0-coincidence (Multiple Optima) . . . . .             | 10         |
| 2.4 Final Optimization Problem . . . . .  | 11         |
| 2.5 Use of Genetic Algorithms to Reach the Optimal Value . . . . .                | 11         |
| 2.6 Statistical Properties of the Linear-Combination Coefficient Estimators . . . | 14         |
| 2.7 Simulation Framework to Validate Results . . . . .                            | 18         |
| <b>3. Results</b>   | <b>21</b>  |
| 3.1 Simulation in 2D . . . . .  | 21         |
| 3.2 Simulation in 3D . . . . .  | 25         |
| 3.3 Multiple Replications of the Simulation Framework . . . . .                   | 27         |
| 3.4 Application to Real Data . . . . .  | 31         |
| <b>Conclusions and Recommendations</b>  | <b>36</b>  |
| <b>Bibliography</b>   | <b>38</b>  |

---

---

# List of Tables

---

---

- 3.1 Performance measures for the single factors (2D) . . . . . 25
- 3.2 Performance measures for the single factors (3D) . . . . . 27
- 3.3 Aggregate performance summary for the replications . . . . . 30
- 3.4 Performance measures for the single factors (Real data) . . . . . 33

---



---

## List of Figures

---



---

|      |  |    |
|------|--|----|
| 2.1  | 0-coincidence p-value function for a two-dimension (2D) factor simulated scenario. . . . .   | 10 |
| 2.2  | 2D illustration of the possible updates of a simplex based on the Nelder-Mead algorithm (Albelwi & Mahmood, 2017). . . . .                                   | 12 |
| 2.3  | Schematic representation of the selection, recombination and mutation steps ( <i>Using Genetic Programming to evolve Trading Strategies</i> , n.d.). . . . . | 13 |
| 2.4  | Procedure to obtain a new vector time series of macroeconomic variables. . .   | 15 |
| 2.5  | Routine to estimate the factors and the coefficients of the coincident index for each iteration of the simulation stage . . . . .                            | 16 |
| 3.1  | Simulation of the state of the economy (2D). . . . .   | 22 |
| 3.2  | Comparison of the simulated versus the estimated factors (2D). . . . .   | 22 |
| 3.3  | Progress Monitoring of the Genetic Algorithms subroutine (2D). . . . .   | 23 |
| 3.4  | Representation of the objective function in polar coordinates (2D). . . . .  | 23 |
| 3.5  | Comparison between the realization of the state of the economy (proxy) and the coincident index (2D). . . . .  | 24 |
| 3.6  | Histogram of the realizations for the optimal value of the objective function.   | 24 |
| 3.7  | Confidence region in Cartesian coordinates (2D). . . . .   | 25 |
| 3.8  | Confidence region in polar coordinates (2D). . . . .   | 26 |
| 3.9  | Simulation of the state of the economy (3D). . . . .   | 27 |
| 3.10 | Comparison of the simulated versus the estimated factors (3D). . . . .   | 28 |
| 3.11 | Representation of the objective function in spherical coordinates (3D). . . . .  | 29 |
| 3.12 | Comparison between the realization of the state of the economy (proxy) and the coincident index (3D). . . . .  | 29 |
| 3.13 | Confidence region in spherical coordinates (3D). . . . .   | 30 |
| 3.14 | Computation of the ISE from Jan-2000 to June-2017. . . . .   | 31 |
| 3.15 | Factors estimated by Nieto et al (2017). . . . .   | 32 |

---

|      |  |    |
|------|--|----|
| 3.16 | Progress of the optimization procedure across the generations (Real data). . | 32 |
| 3.17 | Comparison between the proxy and estimated coincident index (Real data).     | 33 |
| 3.18 | Confidence region in Cartesian coordinates (Real data). . . . .              | 34 |
| 3.19 | Confidence region in polar coordinates (Real data). . . . .                  | 35 |

---

---

## 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 National Product (GNP) is a massive undertaking to measure the overall performance of the economy; however, it is not calculated in a high frequency because it is extremely time-consuming. For that reason, it is necessary to follow alternative approaches to create coincident indices.

During the Twentieth Century, predicting the state of the economy became of capital importance to identify economic growths and decays and to be able to plan accordingly. The first attempts of creating coincident indices were based on the expertise of economic planners, but lacking of any statistical foundation. On the other hand, nowadays, there are solid academic developments that allow conducting the economic index creation based on both economic and statistical knowledge.

This work is the result of an incremental improvement of previous findings developed mainly by Escribano & Peña (1994), Peña & Poncela (2006), Martínez et al. (2016) and Nieto & Chudt (2017). The document is structured as follows: in chapter 1, a concise but comprehensive literature review of the evolution of the different statistical approaches is presented. In chapter 2, the new methodology along with its justification is discussed; while in chapter 3, some simulated and real situations show its performance. Finally, the conclusions, recommendations and the references are listed.

The reader must be aware that the theoretical framework for the analysis presented is the theory of stochastic processes, while the observed and simulated data correspond to time series. For that reason, some of the desirable properties (stationarity, Gaussianity, integration, among others) of the stochastic processes involved are sometimes referred as characteristics of the time series for the sake of simplicity. When this happens, the characteristic mentioned must be understood as a property of the underlying stochastic process where the time series came from.

# CHAPTER 1

---

---

## Literature Review

---

---

According to Stock & Watson (1988), 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, 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 trajectories from each other (with respect to time), there is a latent influence of the state of the economy over all the variables that makes them exhibit common characteristics. For that reason, 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 et al., 1946).

Altissimo et al. (2010) define as main objective for a coincident index to do a valuation about the state of the economy that is:

- *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.
- *Timely*, since it has to provide real-time estimates using all the information at hand.
- *Free from short-run fluctuations*, because the idea is to capture and show the actual trend for the state of the economy without any transient perturbations.

The Gross National Product (GNP) is considered intuitively the best mechanism to track the state of the economy because it is the result of a “census” of the economic activity at a specific time. However, it is not the best choice for a coincident index because (1) it does not provide a real-time measurement of the state of the economy, since it takes a considerable amount of time for the public agencies in charge to compute it and make it available, therefore, it is not available at a high frequency (it is usually

available on a quarterly basis or on a yearly basis); and (2) the GNP is considered to be a simplistic version of the reference cycle since it does not capture all the dynamics in the macroeconomic activity (Stock & Watson, 1989). Those are the main reasons to consider alternative methodologies to create coincident indices based on statistical principles.

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 (1988) were the first ones who proposed the state of the economy as a latent stochastic process that is related in a linear way with each of the indicator variables. The system that they propose follows the expression

$$\Delta Y_t = \beta + \gamma(B) \Delta C_t + u_t, \quad (1.1)$$

being  $\{\Delta Y_t\}$  the vector stochastic process of the first differences of the macroeconomic indicators,  $\beta$  a constant vector,  $\gamma(B)$  a vector depending on the lag operator  $B$ ,  $\{\Delta C_t\}$  the first difference of the latent stochastic process corresponding to the state of the economy, and  $\{u_t\}$  a stochastic process not correlated with  $\{C_t\}$  for any time point. The non-correlation between these two processes has to be intended as an absence of correlation for any pair of random variables,  $C_{t_0}$  and  $u_{s_0}$ , corresponding to any arbitrary time instants  $t_0$  and  $s_0$ , respectively. Additionally, the first difference is considered as a mechanism to guarantee that the stochastic processes involved are stationary since this methodology only deals with stationary stochastic processes.

The representation in (1.1) is consistent with the fact that the indicator variables are composed of: (1) a long-term effect that is a function of the latent state of the economy,  $\{C_t\}$ , and of capital importance for the economic decision makers; and (2) a term that gathers together short-term dynamics and idiosyncratic crashes, hence, with little relevance for economic decision making. An estimation of  $\{C_t\}$  in this model was the first attempt to create an economic index.

Further developments in the area started considering the concept of dynamic common factors 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 et al in (1.1).

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

$$Y_t = P f_t + e_t, t \in \mathbb{Z}, \quad (1.2)$$

where  $\left\{f_t = (f_{1t}, f_{2t}, \dots, f_{rt})^T\right\}$  is denominated the vector of common dynamic factors of the process  $\{Y_t\}$ ,  $T$  is the transpose operator,  $P$  is a matrix of weights of the common dynamic factors with dimension  $m \times r$  and  $\{e_t\}$  is a Gaussian noise process of dimension  $m$ . 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  $\{I_t\}$  that accurately resembles the behavior of  $\{C_t\}$ , i.e., the so-called reference cycle.

In another seminal paper, Stock & Watson (2011) present an overall spectrum of the different alternatives available to estimate dynamic common factors and their chronological appearance in the economic context. They classify the different approaches into 4 branches:

1. *Low-dimension parametric models*: This is the first generation of models according to Stock & Watson (2011). In these models, the factors and the parameters involved in (1.1) are described using a state-space representation and then estimated by means of Gaussian maximum likelihood and the Kalman Filter. The optimality and accuracy of the estimates are based on the verification of the assumptions.

The computational complexity of these models makes them limit the number of parameters and macroeconomic indicators that can be considered. These models will be presented with more detail later on.

2. *High-dimension nonparametric estimation*: This is the second generation of models and it is the type of models that Stock et al formulate (Stock & Watson, 2011). These models require all the series involved to be stationary, but focus their results on asymptotic theory, i.e., on large sample sizes for each series and a large number of series.

These methods use cross-sectional averaging to estimate the factors, which is an analogous technique to Principal Component Analysis (PCA) for multivariate analysis. It has evolved from static PCA through generalized PCA to dynamic PCA. For a further discussion of these models, see Stock & Watson (2002) and Stock & Watson (2011).

Even if it is not mentioned by Stock & Watson (2011) explicitly, there is another approach within this branch that is worth mentioning for the sake of completeness.

Forni et al. (2000) propose a frequency-domain dynamic estimation of the common factors and use this methodology to create a coincident index for the European Union. This approach has the advantage that allows to identify and to isolate the short-term and the long-term effects in the factors and in the resulting coincident index by filtering the corresponding frequencies in the spectrum. For a detailed analysis of this methodology, please refer to Forni et al. (2000), Forni et al. (2005), Cristadoro et al. (2005), and Altissimo et al. (2010).

3. *Nonparametric estimation of state-space models*: These methods combine the approach described in the previous two methodologies trying to keep the advantages (yet also the limitations) of both.
4. *Bayesian models*: The Bayesian models are a parametric alternative to deal with the inconveniences of the Frequentist statistics. This generation developed in parallel with the third generation.

In this brief presentation of the modeling efforts already developed when trying to construct a coincident index, it is notorious that there is a big decision that has to be made (besides the choice among Frequentist and Bayesian statistics) and it regards the requirement of stationarity of the stochastic processes involved. Even if the second generation of models (*the high-dimension nonparametric estimation*) can deal with a large number of

time series and of observations in each time series, it imposes the stationarity as a mandatory condition. On the other hand, the first generation of models (*the low-dimension parametric models*) cannot handle large datasets due to computational complexity, but it does not require the time series to be stationary. Which one of these two characteristics should receive more relevance when constructing a dynamic factor model?

According to Wei (2006) and confirmed by Martínez et al. (2016), when dealing with nonstationary vector time series, differencing allows to make the series under consideration stationary but might eliminate and distort some of the relationships that the series naturally have. For that reason, the methodology followed in this work will consider approaches that allow the macroeconomic indicators to be nonstationary and handle this characteristic in an effective way, as the first generation does.

Following this principle, Martínez et al. (2016) propose a four-step methodology based on dynamic common factors that picks one of the estimated factors as a coincident index. This methodology has the advantage of dealing with nonstationary and cointegrated time series.

Reconsidering the model expressed in (1.2) and assuming that  $\{f_t\}$  follows a  $VARMA(p, q)$  model that looks like

$$\Phi(B) f_t = d + \Theta(B) a_t, \quad (1.3)$$

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 has all its roots outside of the unit circle and do not coincide with any of the roots of  $|\Phi(z)|$ , and the process  $\{a_t\}$  corresponds to a multivariate Gaussian white noise process, independent of the Gaussian white noise process  $\{e_t\}$ ; and whose variance-covariance matrix is of full rank.

Equations (1.2) and (1.3) facilitate to build the state-space representation of the dynamic factor model. 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.

Other criteria that have to be met for the whole state-space model are:

- (a) The elements  $p_{ij}$  from the matrix  $P = (p_{ij})_{i,j}$  must be equal to zero when  $i < j$ , in order for the model to be identifiable.
- (b) All the  $\Phi_i$  and  $\Theta_j$  matrices involved in the operators  $\Phi(B)$  and  $\Theta(B)$  respectively, and the variance-covariance matrices for the error processes,  $\Sigma_e$  and  $\Sigma_a$ , have to be diagonal for the factors to be orthogonal, according to Escribano & Peña (1994).
- (c) The model can have correlated factors, but at least one of the matrices,  $\Sigma_e$  or  $\Sigma_a$ , has to be diagonal for identification purposes.

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 if they have seasonal effects, and to pre-whiten the original data if needed. Pre-whitening is required when the time series of residuals,  $\{\hat{e}_t\}$ , obtained after a first estimation effort; does not seem to be the realization of a multivariate white noise process and there are no specification errors in the model.

Nieto et al. (2016) propose an alternative methodology to deal with seasonality when estimating dynamic common factor 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)

$$C_y(k) = \frac{1}{S^{2d+1}} \sum_{t=k+1}^S (y_{t-k} - \bar{y})(y_t - \bar{y})^T, k = 0, 1, 2, \dots, \quad (1.4)$$

and the canonical correlation matrices (CCM)

$$\hat{M}_1(k) = \left[ \sum_{t=k+1}^S y_t y_t^T \right]^{-1} \left[ \sum_{t=k+1}^S y_t y_{t-k}^T \right] \left[ \sum_{t=k+1}^S y_{t-k} y_{t-k}^T \right]^{-1} \left[ \sum_{t=k+1}^S y_{t-k} y_t^T \right], k = 0, 1, 2, \dots, \quad (1.5)$$

where  $S$  is the sample size of the vector time series,  $d$  is the order of integration of the vector  $\{Y_t\}$  and  $k$  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.

There are two variations of this approach that should be mentioned. First, Lam et al. (2012) extend this approach when there is a high-dimensional vector of time series. In addition, Bujosa et al. (2013) point out that Peña and Poncela's approach relies on the assumption that the error processes are Gaussian, zero-mean and full rank variance-covariance matrix white noise processes, which is hardly verified in reality. For that reason, they rather use an exploratory approach on the eigenvalues and eigenvectors of the SGCV to identify if there are relatively large eigenvalues (in absolute value) associated to stable eigenvectors.

In this work, for the estimation of the factors, the MARSS package in R® will be used (Holmes et al., 2012).

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 to assess the adequacy of each factor as 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 a result of a nonparametric test of  $i$ -coincidence,  $i \in \mathbb{Z}$ .

For a formal presentation of the topic, please refer to Nieto & Chudt (2017).

As it has been mentioned before, the state of the economy is a latent stochastic process, therefore, its realizations are not observable. For that reason, it is necessary to identify a good proxy of the state of the economy,  $\{\hat{c}_t\}$ , to compare with. In most of the cases, a high-frequency interpolated series of the GNP is a reasonable starting point.

Nieto & Chudt (2017) also mention that the coincident profile 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. For that reason, they apply the coincident profile procedure to the first difference of the proxy and each of the factors arguing that if two functions have the same derivative in calculus, they are identical up to a 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 then, analyze the direct implications in the real context of the problem.

In the same work, Martínez et al. (2016) pose for a future work the possibility of constructing a coincident index as a linear combination of the common factors and assessing its statistical adequacy to predict the state of the economy.

---



---

## Methodology

---



---

### 2.1 Formulation of the Problem

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

This work intends to explore among all the possible linear combinations of factors to identify a particular combination that maximizes a measure of goodness of fit. In this case, this measure is related to the coincident profile, but requires other elements to work properly.

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

$$I_t := \bar{\alpha}^T f_t = \sum_{i=1}^r \alpha_i f_{it}, t \in \mathbb{Z}, \quad (2.1)$$

being  $\alpha_i \in \mathbb{R}, i = 1, 2, \dots, r$ .

The coincident profile proposed by Martínez et al. (2016), and then refined by Nieto & Chudt (2017), is the presentation of the p-value for several tests of  $i$ -coincidence, with  $i$  varying from -3 to 3. For each one of the estimated factors, the level of  $i$ -coincidence is determined by choosing the value of  $i$  for which the p-value for  $i$ -coincidence is the maximum among the coincident profile and is greater than a pre-specified significance level.

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 leads the researchers to choose sometimes as coincident index a factor with a coincidence level other than 0-coincidence (but fairly close) if none of the factors exhibits 0-coincidence.

However, the attention of this work will be focused on the statistical test to measure 0-coincidence because, within the coincident profile, it is the only test that measures directly the ability of an index to be coincident.

Given that the 0-coincidence is based on a statistical test that measures the coincidence between the turning points of two series (Banerji, 1999), it is invariant under transformations of scale, thus any nonnegative scaling 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),  $\vec{\alpha}^T f$  and  $k\vec{\alpha}^T f$  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.,  $\vec{\alpha}$  is such that

$$\|\vec{\alpha}\|_2^2 = \sum_{i=1}^r \alpha_i^2 = 1. \quad (2.2)$$

The feasible region is defined in terms of an equality and not as  $\|\vec{\alpha}\|_2^2 \leq 1$  because, even if this region 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 for every  $k \in (0, 1)$ .

Additionally, the  $L_2$  norm was preferred over  $L_1$  norm for instance, because  $L_1$  usually is employed for variable (or factor in this case) selection in regularization scenarios, which is undesirable if the idea is to look for nontrivial linear combinations of the factors. Besides, the convexification for a  $L_1$  norm restriction with equality is way more complicated.

Summarizing what has been discussed so far, a first attempt to address the problem of finding a coincident index by means of linear combinations of the common dynamic factors would be to solve the following maximization problem

$$\begin{aligned} \max_{\vec{\alpha}} p - value_0(\{\Delta \hat{c}_t\}, \{\Delta i_t\}), \\ \text{s.t. } \|\vec{\alpha}\|_2^2 = 1, \end{aligned} \quad (2.3)$$

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  $p - value_0(\cdot, \cdot)$  the function that calculates the p-value for the 0-coincidence between  $\{\hat{c}_t\}$  and  $\{i_t\}$ . The acronym "s.t." stands for "subject to:" and precedes the constraints in the optimization problem. The difference operator is introduced following Nieto & Chudt (2017).

Weierstrass' theorem for optimization problems guarantees that if the feasible region is a compact set (which indeed is in this case) and the objective function is continuous, the optimal value is achieved within the feasible region (Bazaraa et al., 2013). However, as it will be seen in one of the following subsections, the  $p - value_0(\cdot, \cdot)$  function has multiple optima and that does not allow achieving a unique global maximum.

It is also important to highlight that there is no closed-form expression to evaluate  $p - value_0(\cdot, \cdot)$  given any two arguments. This function has to be evaluated by means of a permutation test, as it can be consulted in Nieto & Chudt (2017). This feature leaves out of consideration any derivative-based optimization algorithm.

For the sake of simplifying the notation further ahead, let

$$p(\Delta\hat{c}_t, \Delta i_t) := p - \text{value}_0(\{\Delta\hat{c}_t\}, \{\Delta i_t\}). \quad (2.4)$$

## 2.2 Use of Spherical Coordinates

Another big challenge of this research effort was to find a valid framework to perform the optimization above-mentioned given its particular characteristics. Two of the challenges were already mentioned: the intuitive objective function for this problem seems to have multiple optima, and there is no closed-form expression to use derivative-based methods because mathematical properties are not guaranteed. Besides that, the constraint imposed to the coefficients for the factors in (2.2) defines a non-convex space in  $\mathbb{R}^r$ , which makes it almost intractable for any of the optimization algorithms that are usually implemented.

To overcome this challenge, the problem was solved using spherical coordinates, since under that representation, the feasible set becomes convex. Recall that for any point in the Cartesian coordinates system for  $\mathbb{R}^r : (\alpha_1, \alpha_2, \dots, \alpha_r)$ , there is an equivalent representation in the spherical coordinates system:  $(\rho, \theta, \phi_1, \phi_2, \dots, \phi_{r-2})$  where  $\rho \in \mathbb{R}_+$ ,  $\theta \in [0, 2\pi)$  and  $\phi_i \in [0, \pi)$ ,  $i \in \{1, 2, \dots, r-2\}$ . The equations that describe the relationship between these two representations, according to Blumenson (1960), are

$$\begin{aligned} \alpha_r &= \rho \cos(\phi_{r-2}), \\ \alpha_{r-1} &= \rho \sin(\phi_{r-2}) \cos(\phi_{r-3}), \\ \alpha_{r-2} &= \rho \sin(\phi_{r-2}) \sin(\phi_{r-3}) \cos(\phi_{r-4}), \\ &\vdots \\ \alpha_2 &= \rho \sin(\phi_{r-2}) \sin(\phi_{r-3}) \sin(\phi_{r-4}) \sin(\phi_{r-5}) \dots \sin(\phi_1) \sin(\theta), \\ \alpha_1 &= \rho \sin(\phi_{r-2}) \sin(\phi_{r-3}) \sin(\phi_{r-4}) \sin(\phi_{r-5}) \dots \sin(\phi_1) \cos(\theta). \end{aligned} \quad (2.5)$$

The constraint imposed to the coefficients of the linear combination presented in (2.2) is equivalent to the surface of a hypersphere in  $\mathbb{R}^r$ , therefore, its representation in spherical coordinates is quite simple

$$\begin{aligned} \rho &= 1, \\ 0 &\leq \phi_i \leq \pi, i \in \{1, 2, \dots, r-2\}, \\ 0 &\leq \theta \leq 2\pi. \end{aligned} \quad (2.6)$$

The representation in (2.6) corresponds to a nice polyhedral region that is convex. It is also important to notice that since  $\rho = 1$ , this parameter actually does not vary and the feasible region in spherical coordinates can be seen as a subset of  $\mathbb{R}^{r-1}$  whose variables are  $(\theta, \phi_1, \phi_2, \dots, \phi_{r-2})$ .

The optimization problem in (2.3) can now be expressed as

$$\begin{aligned} \max_{(\theta, \phi_1, \phi_2, \dots, \phi_{r-2})} & p(\Delta\hat{c}_t, \Delta i_t), \\ \text{s.t. } i_t &= \sum_{i=1}^r \alpha_i(\theta, \phi_1, \phi_2, \dots, \phi_{r-2}) \cdot \hat{f}_{it}, \\ & 0 \leq \phi_i \leq \pi, i \in \{1, 2, \dots, r-2\}, \\ & 0 \leq \theta \leq 2\pi. \end{aligned} \quad (2.7)$$

Once the optimal solution for this problem is computed, the original coefficients can be calculated by replacing the optimal values for (2.7) into the set of equations in (2.5) and recalling that  $\rho = 1$ .

### 2.3 Inconveniences with the 0-coincidence (Multiple Optima)

As it has been mentioned before, the function  $p(\cdot, \cdot)$  may have multiple optima (it might reach the value of 1 for several combinations of the factors). Figure 2.1 shows a plot of this function in the case of only two factors (which allows to represent the problem using only one variable in spherical coordinates).

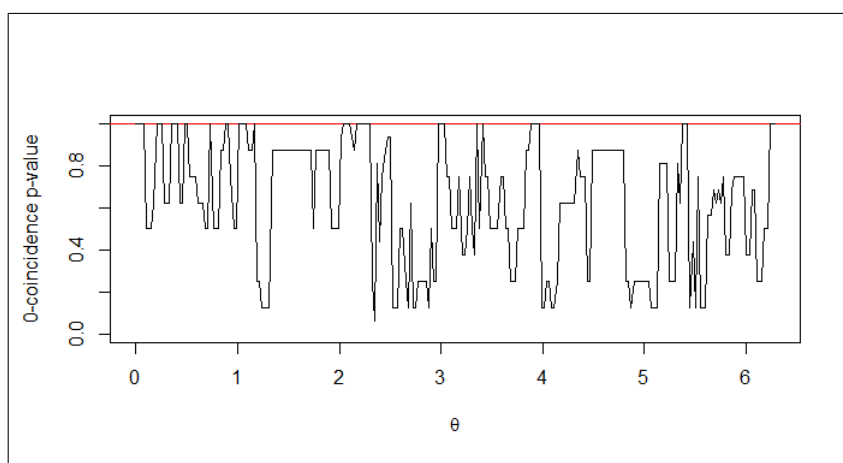


FIGURE 2.1. 0-coincidence p-value function for a two-dimension (2D) factor simulated scenario.

*The red line represents the maximum p-value (equal to 1) that is actually attained over the search spectrum. It is possible to see how several points attain this value.*

This problem had not been faced by previous research efforts in the area, such as Martínez et al. (2016) and Nieto & Chudt (2017), since their approach was to pick among a finite set of alternatives (the set of estimated factors) a coincident index, instead of exploring over an uncountable set.

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, and based also on the fact that Nieto & Chudt (2017) apply the first differences to the time series in order to measure the 0-coincidence, a good candidate to be part of the objective function is the cross-correlation at lag 0 between  $\{\Delta\hat{c}_t\}$  and  $\{\Delta i_t\}$ ,  $cor(\Delta\hat{c}_t, \Delta i_t)$ , assuming that the bivariate process  $\{(\Delta\hat{C}_t, \Delta I_t)\}$  is stationary.

The cross-correlation function allows considering 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. Nieto & Chudt (2017) use a similar

approach to select the original set of variables to be included in the analysis of their paper.

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  $I(1)$ . 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 it is 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 and overshadows the other component.

The two components will be included in the objective as the terms of a simple sum since, under this condition, the whole objective function varies from -1 to 2 and it is possible to see how far a given value is from the boundaries.

Some sort of weighted average might be considered for future work.

## 2.4 Final Optimization Problem

By taking into account all the elements described in the previous sections, the definitive optimization problem to be solved is

$$\begin{aligned} \max_{(\theta, \phi_1, \phi_2, \dots, \phi_{r-2})} z &= [p(\Delta\hat{c}_t, \Delta i_t) + cor(\Delta\hat{c}_t, \Delta i_t)], \\ \text{s.t. } i_t &= \sum_{i=1}^r \alpha_i(\theta, \phi_1, \phi_2, \dots, \phi_{r-2}) \cdot \hat{f}_{it}, \\ 0 &\leq \phi_i \leq \pi, i \in \{1, 2, \dots, r-2\}, \\ 0 &\leq \theta \leq 2\pi. \end{aligned} \tag{2.8}$$

This is a constrained optimization problem with an apparent non-convex objective function (based on the plots generated for some cases) and a convex feasible region.

## 2.5 Use of Genetic Algorithms to Reach the Optimal Value

As it has been mentioned previously, the objective function of the problem stated in (2.8) does not have an analytical expression to be computed. Its calculation depends on the estimation of a proportion based on all the possible permutations, as it is described in Nieto & Chudt (2017).

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 require only the availability of an objective function are denominated derivative-free algorithms.

Rios & Sahinidis (2013) make a comprehensive, yet brief, description of the types of algorithms that can be implemented for derivative-free optimization efforts. The main three classifications are:

1. *Based on the search scope (local/global)*: Local algorithms improve the solution around a given initial condition, therefore, they are very sensitive to this initial condition. On the other hand, the global algorithms tend to expand the search space and reach a global optimum (if possible).
2. *Based on the type of approach (direct/model-based)*: Direct search algorithms exploit only the evaluation of the objective at different values of the input variables to define how to improve the search, while model-based algorithms use a surrogate or an assisting function, related to the original objective, to improve the search.
3. *Based on the consideration of randomness (stochastic/deterministic)*: Stochastic algorithms involve random search steps when they switch from one iteration to another, while deterministic algorithms do not use any type of randomness.

Given the characteristics of the problem under analysis and the particularities of its objective function, only two algorithms were considered: The Nelder-Mead algorithm and Genetic Algorithms. Both were available in the software R®, which enhanced the connection of this subroutine with the whole methodology described.

**The Nelder-Mead algorithm (Nelder & Mead, 1965)**: This algorithm was proposed by J. A. Nelder and R. Mead in 1965. It is a local, direct-search oriented, deterministic algorithm and it is based on the idea of updating the limit points of a simplex (which is basically a polyhedral subset inside the feasible region) by replacing the corner with the worst objective value. The update procedure can only be done by following specific transformations of the simplex: reflection, expansion, inside- and outside- contractions, and shrinkage (see Figure 2.2). Even if it is one of the first methods ever proposed, it is still widely used due to its simplicity, flexibility, and reliability. Convergence to stationary points has been proven in some scenarios, but it might produce misleading results, as discussed by Rios & Sahinidis (2013).

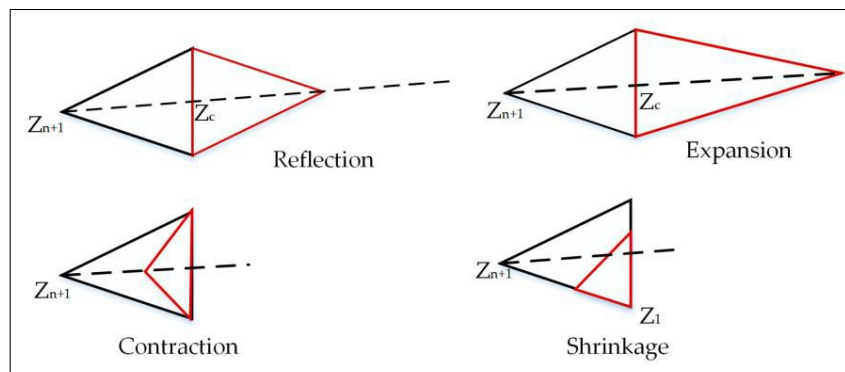


FIGURE 2.2. 2D illustration of the possible updates of a simplex based on the Nelder-Mead algorithm (Albelwi & Mahmood, 2017).

**Genetic Algorithms (Sastry et al., 2014)**: This algorithm was first introduced by J.H. Holland in 1975. It is a global, direct-search, stochastic algorithm inspired by the

rules of evolution and chromosomic interactions. The idea of the algorithm is to make sure that the fittest individuals in the population survive after each generation of the model. It is based on the following steps:

1. Initialization: An initial population of candidate solutions is selected from the feasible region. This process can be done in a totally random way or by introducing some previous information of the problem.
2. Evaluation: For all the current members of the population, the fit function is evaluated. This function is usually the objective function of the problem.
3. Selection: Considering one among a wide list of criteria, the fittest individuals in the population are selected to be the parents of the following generation.
4. Recombination: The information that the selected parents carry is shared and combined to create new genetic profiles for the offspring. It is analogous to the crossover that chromosomes perform during the cell division phases.
5. Mutation: Based on a random mechanism, the genetic profiles for the offspring are altered to give chance to accidental and arbitrary events to maybe come up with a better solution.
6. Replacement: The new offspring replaces the old generation in a partial or in a total way depending on the characteristics of the algorithm.

The steps 2 to 6 are repeated until some convergence criterion is met. In most of the cases, the criterion is associated with a predefined number of maximum generations or with the proximity of the best values between successive generations.

As it is mentioned by Rios & Sahinidis (2013), in 1978, J.D. Bethke adapted the genetic algorithms to deal with continuous variables by expressing their values in binary code and considering each position with a zero or a one as one gene (Bethke, 1978). Figure 2.3 shows schematically how steps two, three and four are conducted.

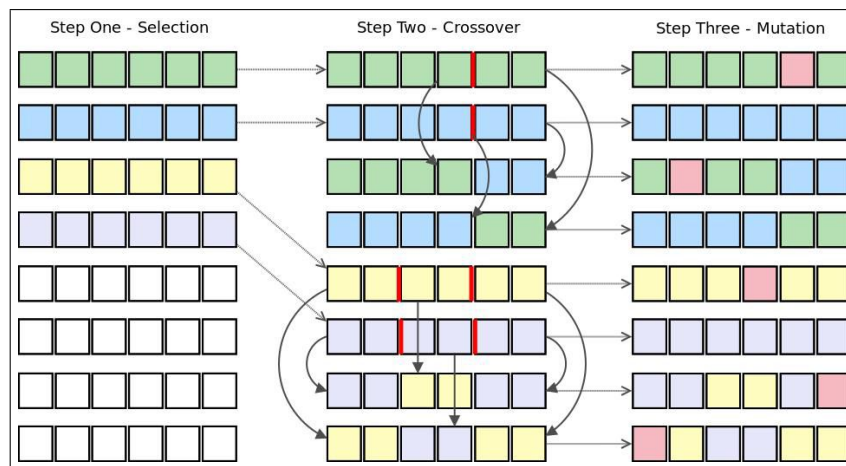


FIGURE 2.3. Schematic representation of the selection, recombination and mutation steps (*Using Genetic Programming to evolve Trading Strategies*, n.d.).

Both algorithms were tested with simulated results, although, as it was already expected because of the characteristics of each algorithm, the genetic algorithms outperformed the Nelder-Mead algorithm in attaining a better optimal point. For that reason, in future sections, the results presented were obtained using an adaptation of genetic algorithms in R® by means of the package GA (Scrucca et al., 2013).

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

## 2.6 Statistical Properties of the Linear-Combination Coefficient Estimators

The estimators of the unknown coefficients  $\alpha_i \in \mathbb{R}, i = 1, 2, \dots, r$ ; are such that

$$\hat{\alpha}_i := \alpha_i \left( \hat{\theta}, \hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_{r-2} \right), \quad (2.9)$$

where:

$$\left( \hat{\theta}, \hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_{r-2} \right) = \underset{(\theta, \phi_1, \phi_2, \dots, \phi_{r-2})}{\arg \max} \left[ p \left( \Delta \hat{C}_t, \Delta I_t \right) + cor \left( \Delta \hat{C}_t, \Delta I_t \right) \right], \quad (2.10)$$

subject to the constraints presented in (2.8).

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 (2.10).

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 statistical inference for this particular problem.

So far, it is known that these point estimates come from an optimal procedure that attempts to maximize the level of 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 dynamic common factors and the coincident index frameworks apply.

However, it is important to consider mechanisms to construct confidence regions or hypothesis tests 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 efforts, like Martínez et al. (2016) and Nieto & Chudt (2017).

For instance, assume that some estimates for the coefficients 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 common factors is

the best coincident index that can be created. That scenario can be translated into the following hypothesis system for some  $i, i = 1, 2, \dots, r$

$$\begin{cases} H_0 : (\alpha_1, \dots, \alpha_i, \dots, \alpha_r)^T = (0, \dots, 1, \dots, 0)^T \\ \text{v.s.} \\ H_1 : (\alpha_1, \dots, \alpha_i, \dots, \alpha_r)^T \neq (0, \dots, 1, \dots, 0)^T \end{cases} \quad (2.11)$$

The null hypothesis in system (2.11) supports the idea that only factor  $i$ , for a particular  $i$ , 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  $i, i \in \{1, 2, \dots, r\}$ ; there is evidence to conclude that the coincident index cannot be composed of only one factor and that result would play in favor of this general procedure.

The question that arises at this point is: how can system (2.11) be assessed based on the nature of the estimation procedure presented here?

A theoretical answer to this question is complicated and cumbersome, but a simulation-based approach seems to be a reasonable alternative. The methodology presented here can be easily extended to more general hypothesis systems.

Following the general representation of the common dynamic factors model, and assuming that for a given vector of macroeconomic time series  $\{y_t\}$ , the common dynamic factors,  $\{\hat{f}_t\}$ , the matrices  $\hat{\Sigma}_a$ ,  $\hat{\Sigma}_e$  and  $\hat{P}$  have been estimated; and also that the corresponding evolution dynamics for the dynamic factors, i.e., their VARIMA structure, has been identified; it is possible to simulate new results from that system and generate different realizations of the coefficients by following the steps in Figure 2.4 and in Figure 2.5.

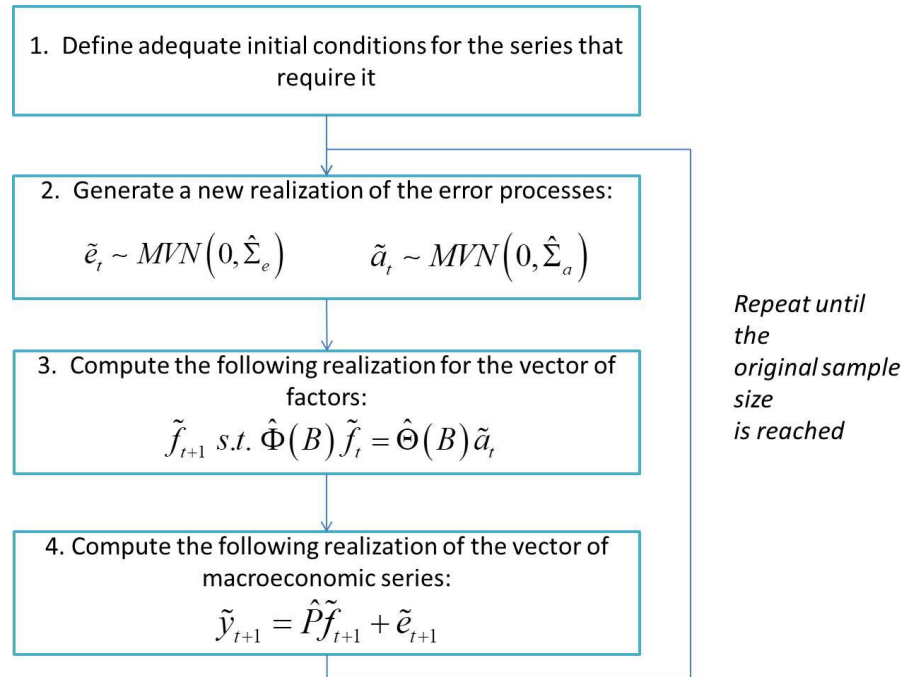


FIGURE 2.4. Procedure to obtain a new vector time series of macroeconomic variables.

It is important to notice that each of the simulated vector time series  $\{\tilde{y}_t\}^{(j)}, j \in \{1, 2, \dots, N\}$  for some simulation sample size  $N$ , is generated using the same parameters that were estimated in the previous stage of the analysis. In case the normality assumption for the multivariate white noise processes does not hold, the step 2 in 2.4 can be replaced for a nonparametric re-sampling step based on the residuals obtained during the parameter estimation procedure.

Once a new multivariate time series,  $\{\tilde{y}_t\}^{(j)}$ , has been simulated according to the state-space layout previously estimated, it is necessary to repeat the procedure of estimation of the dynamic common factors and then, identify the optimal combination of factors.

Figure 2.5 presents the procedure followed. The term  $(\hat{p} + \widehat{cor})^{(j)}$  denotes the estimate of the objective function in each draw of the  $N$  simulations.

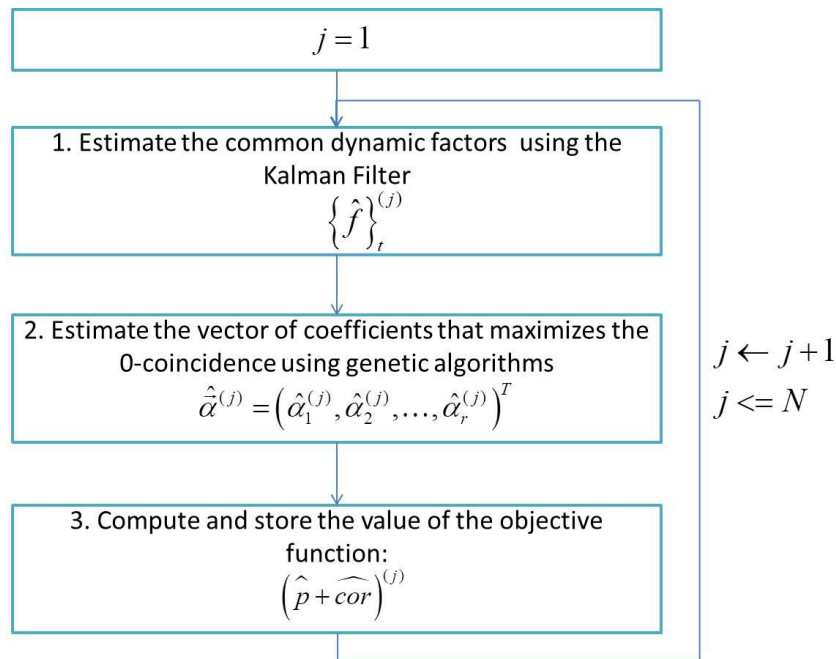


FIGURE 2.5. Routine to estimate the factors and the coefficients of the coincident index for each iteration of the simulation stage

It would seem intuitive to use the whole set of estimates for the vector of coefficients,  $\{\hat{\alpha}^{(j)}\}_{j=1}^N$ , to draw inferences about the real coefficients. For example, the empirical distribution of this sample could suggest a distributional form for the actual estimator and, based on a confidence level and the quantiles of that distribution, a confidence region for the real parameters could be established. The type of hypothesis formulated in (2.11) could be easily assessed by simply determining if the value under the null hypothesis is contained in such a confidence region, for a given confidence (or significance) level.

Nevertheless, the approach mentioned above would show with no doubt misleading conclusions in terms of inference about the real parameters. The reason behind this fact lies in the way the simulation is conducted and the properties of the common dynamic factor models.

There is an analogy between common dynamic factor models and the principal component analysis in multivariate statistics. Recall that the principal components analysis aims at reducing the dimension of a multivariate random vector by creating orthogonal linear combinations of the original variables that maximize their own variance (Johnson & Wichern, 2002). In essence, to compute the principal components of a random vector, it suffices to compute the eigenvectors and the eigenvalues of the variance-covariance matrix of the vector and then, the eigenvector with the highest eigenvalue will correspond to the first principal component (the one with the highest variance), the second highest eigenvalue to the second principal component, and so on.

However, it is important to remember that 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 common dynamic factors models, the factors are analogous to the principal components of a multivariate random vector. For that reason, in their estimation procedure, sometimes it might result an estimation similar to  $f_{it}$  and in other cases similar to  $-f_{it}$  for some  $i$ , because both capture the same variability.

Additionally, in the case of the classic principal component analysis, the sorted eigenvalues can help to identify the natural order of the principal components. In the case of the common dynamic factor models, there is not a unified criterion to sort the factors and, in some estimation efforts, the factor that came first in a previous iteration can appear in a different position.

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

Regardless of the ordering or the inverted sign of some factors, the objective function  $(\hat{p} + \widehat{cor})^{(j)}$  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  $\left\{ (\hat{p} + \widehat{cor})^{(j)} \right\}_{j=1}^N$  can be considered as a realization of a random sample of the corresponding random variable (which is a function of the underlying stochastic processes).

Bickel & Doksum (2015) express that, given a confidence interval  $C(\mathbf{X})$  for a parameter  $\theta$  with a given level of confidence  $1 - \delta$ ,  $\delta \in (0, 1)$ ; a confidence interval with the same level of confidence for a function of the parameter  $q(\theta)$ , can be defined as

$$q(C(\mathbf{X})) := \{q(\theta), \theta \in C(\mathbf{X})\}. \quad (2.12)$$

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

An analogous result will be applied to the problem to be able to find a confidence region for the coefficients based on a confidence interval for the objective function. The interval considered for the objective function will be a unilateral confidence interval for two reasons:

- 1) The objective function is bounded. It has an upper bound of 2.

- 2) Since the main objective of the optimization is to maximize the objective function, there is no reason to exclude values in the right tail (close to the upper bound).

Considering that an interval for the objective function  $\left[ p \left( \Delta \hat{C}_t, \Delta I_t \right) + cor \left( \Delta \hat{C}_t, \Delta I_t \right) \right]$ , which is a parameter, has the general form  $[L, 2]$  and considering that there exists a function  $g(\cdot)$  such that

$$g(\vec{\alpha}) = p \left( \Delta \hat{C}_t, \Delta I_t \right) + cor \left( \Delta \hat{C}_t, \Delta I_t \right), \quad (2.13)$$

a confidence region of level  $1 - \delta$  for the vector of coefficients  $\vec{\alpha}$  corresponds to the set

$$\left\{ \vec{\alpha} : g(\vec{\alpha}) \in [L, 2] \text{ and } \|\vec{\alpha}\|_2^2 = 1 \right\}. \quad (2.14)$$

If  $z^*$  is defined as  $g(\vec{\alpha})$ , the following equivalences hold in the common probability space  $(\Omega, \mathfrak{S}, P)$  that allows all the random applications involved to be proper random variables

$$\begin{aligned} 1 - \delta &= P \{ \omega \in \Omega : z^* \in [L(\omega), 2] \} \\ &= P \{ \omega \in \Omega : g(\vec{\alpha}) \in [L(\omega), 2] \} \\ &= P \{ \omega \in \Omega : \vec{\alpha} \in g^{-1}([L(\omega), 2]) \}, \end{aligned} \quad (2.15)$$

where  $g^{-1}([L(\omega), 2])$  is the set of pre-images in  $\mathbb{R}^r$  that under  $g(\cdot)$  fall within  $[L, 2]$ . Notice that  $g(\cdot)$  does not need to be invertible for the set  $g^{-1}([L(\omega), 2])$  to be well-defined (Bloch, 2011).

In the application of this result, there is no analytical expression for the function  $g(\cdot)$ , for that reason, 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. The application of these results will be presented in the next section.

## 2.7 Simulation Framework to Validate Results

To check the validity of the methodology proposed in this work, some simulation results will be presented. This section is very important because it highlights some of the particular features of the general problem that are trying to be solved and allows evidencing how the methodology can be implemented.

### Simulation of a Base Scenario

The simulation framework of a base condition will be motivated by Stock et al's approach (Stock & Watson, 1988), in the sense that the state of the economy is the main common feature among the macroeconomic variables, as it is shown in equation (1.1).

The state of the economy,  $\{C_t\}$ , will be assumed to follow a general ARIMA model, i.e.,

$$\{C_t\} \sim ARIMA(p, d, q). \quad (2.16)$$

According to Stock et al's model, this factor is the input to generate the vector of stochastic processes for the macroeconomic variables and there is not an intermediate connection with a vector of factors (Stock & Watson, 1988). However, for the sake of the analysis and to be able to fully implement the methodology previously described, it will be assumed that the state of the economy generates a vector of  $r$  common factors via the expression

$$f_t = P_f C_t + u_t, \quad (2.17)$$

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

$$Y_t = P f_t + v_t, \quad (2.18)$$

where  $v_t \sim MVN(0, \Sigma_v)$  and  $\Sigma_v$  will be assumed to be

$$\Sigma_v = \text{diag} \left\{ i^2/t(n) : 1 \leq i \leq n \right\}. \quad (2.19)$$

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

As it was stated before, the intermediate step of calculating the factors was included to be able to compare the original simulated results 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 for the vector of factors is used as an input in the simulation of the framework. However, this does not necessarily imply that the variables will 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 common factors present in each scenario can vary from 1 to that dimension size.

Once a full set of realizations for the state of the economy, the vector of dynamic common factors and the vector of macroeconomic variables has been obtained, the methodology presented in section 3 can be applied. From now on, it will be assumed that only the realizations of the vector of macroeconomic variables are available. The realizations for the factors and the state of the economy will only be presented to compare with the estimated results.

### Estimation of the Common Dynamic Factors

Firstly, it is necessary to estimate the amount and the realizations of the common dynamic factors following Peña & Poncela (2006). Since the objective of this work is not to deal with the issues in the identification of the number of factors, it will only be assumed that the number of factors,  $r$ , is correctly determined, thus, it is equal to the real one that was used in the simulation. Although 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 introduce variability in the results, it will always be assumed that the factors follow a multivariate random walk dynamics

$$f_t = f_{t-1} + a_t, \quad (2.20)$$

where  $a_t \sim MVN(0, \Sigma_a)$  and  $\Sigma_a$  is a diagonal matrix.

The random walk assumption is fairly common in the state-space models because even if it is a simple structure that does not depend on parameters, it allows to capture fairly general evolution dynamics as Hamilton (1994) and Pivetta & Reis (2007) suggest.

It must be mentioned that all the other conditions necessary for the state-space model presented in chapter 1 to be identifiable were considered.

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 contemporary cross-correlation with the realization of the proxy for the state of the economy.

In this work, it will be assumed that the proxy available is the realization of the state of the economy itself. However, this will not be possible in real life, since the state of the economy is a latent stochastic process.

### **Simulation Sampling and Inference about the Coefficients**

Following the procedures presented in Figure 2.4 and Figure 2.5, it is possible to conduct inference about the coefficients, i.e., create a confidence region with a certain level of confidence and to assess certain systems of hypothesis.

---

## Results

---

This section presents the results of a couple of simulations and an application to real data to show how the methodology already described can be applied. The first case is a 2D scenario (two-dimension example in the space of the factors) while the second one is a 3D (three-dimension example) scenario. Finally, the application was made to the results previously obtained by Nieto & Chudt (2017).

### 3.1 Simulation in 2D

For this simulation, it was assumed that the number of common factors 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 but considering the conditions previously stated. For this particular scenario, the matrices were:

$$P_f = \begin{bmatrix} 0.18 \\ 0.49 \end{bmatrix} P = \begin{bmatrix} 0.49 & -0.98 & -6.03 & -0.91 & 3.53 \\ 0 & -2.15 & -1.39 & 2.87 & -5.22 \end{bmatrix}^T \quad (3.1)$$

Once the state of the economy, the factors and the macroeconomic variables have been simulated, the realization of the macroeconomic variables were used to estimate the 2 factors using the Kalman Filter. The results for the simulations and the estimated factors are presented in Figure 3.1 and Figure 3.2.

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. This is one of the main motivations to actually include the possibility of linear combinations (in some cases with negative coefficients) to build a coincident index.

Figure 3.3 shows the overall progress through the generations of the optimization subroutine based on genetic algorithms. Even if there is not a certain criterion of convergence for the genetic algorithms 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

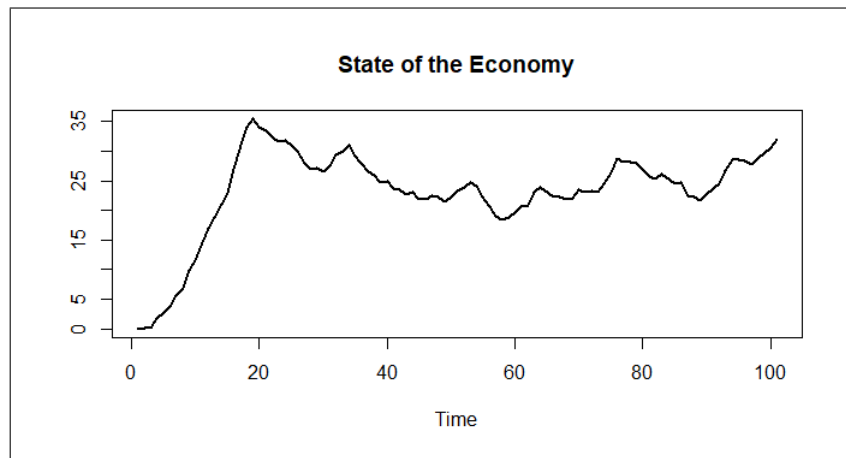


FIGURE 3.1. Simulation of the state of the economy (2D).

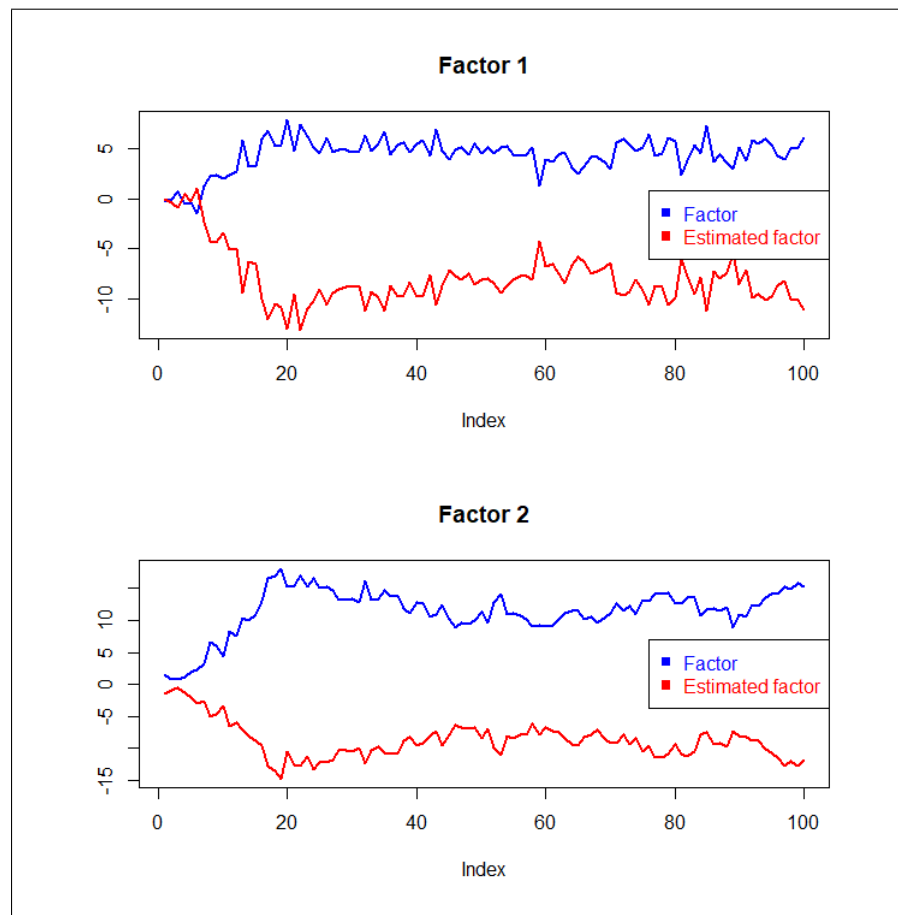


FIGURE 3.2. Comparison of the simulated versus the estimated factors (2D).

the best value, the median and the mean values along the generations seem to be close together are good signs of the quality of the solution in terms of optimality.

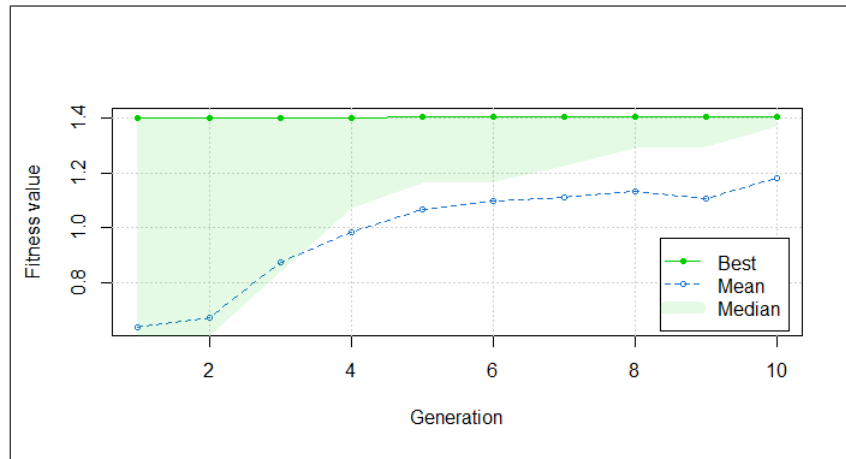


FIGURE 3.3. Progress Monitoring of the Genetic Algorithms subroutine (2D).

For each generation, the best, the median and the mean value of the individuals of that generation are presented.

The results showed that the optimal value for this case was  $z^* = 1.405$ , corresponding to a 0-coincidence p-value of 1 and a correlation of 0.405. This optimal value is attained at  $\theta^* = 4.0214$  in polar coordinates, and  $\alpha_1^* = -0.637$ ,  $\alpha_2^* = -0.771$  in Cartesian coordinates. According to this result, the coincident index is close to an average of the negatives of the two estimated factors. Figure 3.4 shows the objective function over the whole spectrum and Figure 3.5 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.

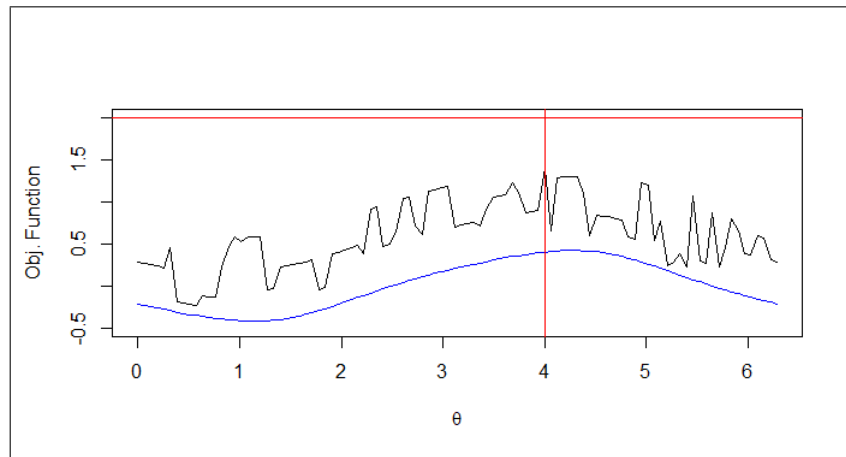


FIGURE 3.4. Representation of the objective function in polar coordinates (2D).

The optimal value is represented by a red vertical line, while the correlation function is represented by the blue curve.

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

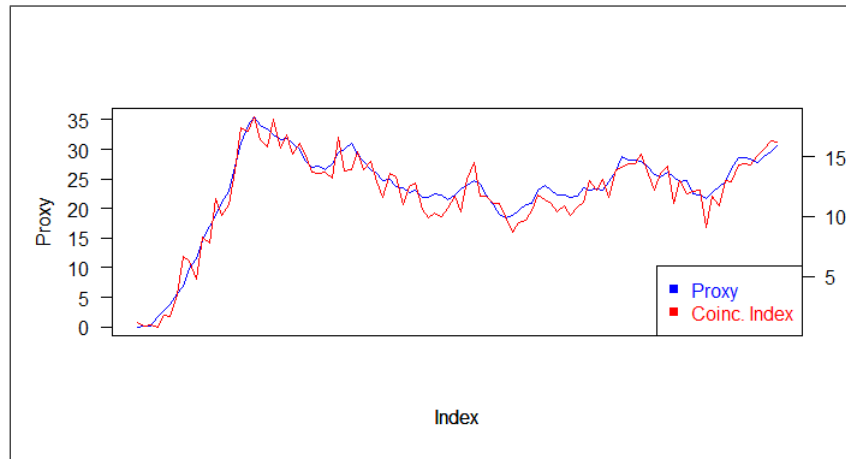


FIGURE 3.5. Comparison between the realization of the state of the economy (proxy) and the coincident index (2D).

optimal value of the objective function is  $[0.954, 2]$ . The histogram for 20 realizations of the objective function is presented in Figure 3.6. The sample size can be increased in case more accuracy is required, however, only 20 realizations were considered in this example because of computation time and because the main objective of this section is to illustrate the use of the methodology.

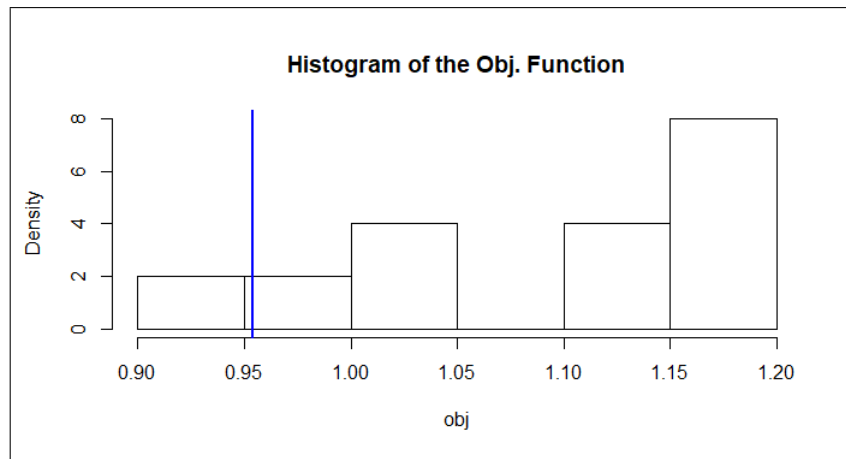


FIGURE 3.6. Histogram of the realizations for the optimal value of the objective function.

*The value represented by the blue line corresponds to the 5% percentile.*

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 polar coordinates. Figure 3.7 and Figure 3.8 show the confidence region in Cartesian and in polar coordinates.

According to the results of the confidence regions, it can be concluded that no single factor can actually be a coincident index. Their corresponding performance measures are presented in Table 3.1.

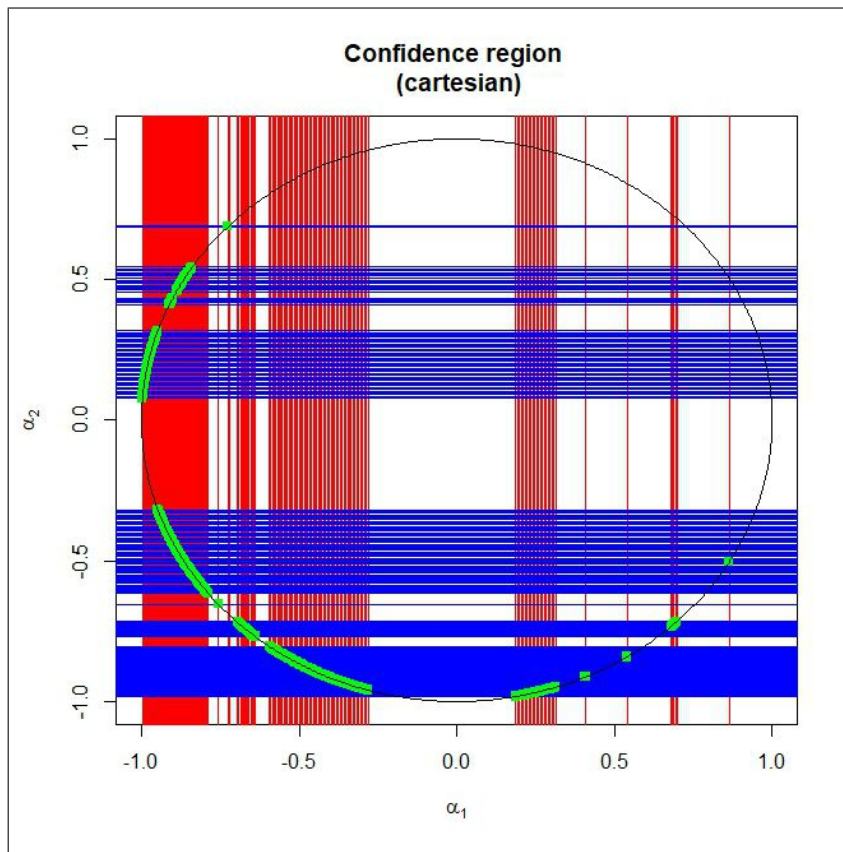


FIGURE 3.7. Confidence region in Cartesian coordinates (2D).

The green points represent the values over the unit circle that lie within the confidence region while the red and the blue lines represent the projections of these points on each of the axes.

TABLE 3.1. Performance measures for the single factors (2D)

| Performance measure | Factor 1 | Factor 2 |
|---------------------|----------|----------|
| $z^*$               | 0.286    | 0.261    |
| $p$ -value          | 0.5      | 0.625    |
| $cor$               | -0.214   | -0.364   |

## 3.2 Simulation in 3D

For this simulation, it was assumed that the number of common factors 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  $\phi = 0.7$ .

The matrices  $P_f$  and  $P$  were generated by a random mechanism but considering the conditions previously stated. For this particular scenario, the matrices were:

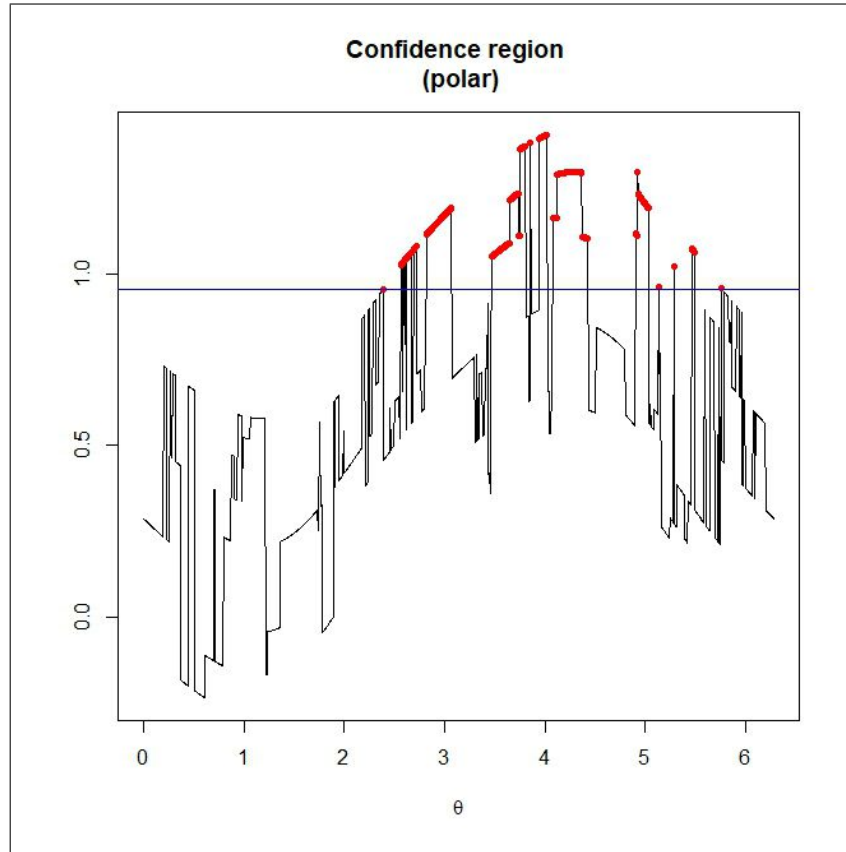


FIGURE 3.8. Confidence region in polar coordinates (2D).

*The x-coordinate of the red dots corresponds to the confidence region in polar coordinates. The horizontal blue line represents the lower limit of the confidence interval for the objective's optimum.*

$$P_f = \begin{bmatrix} -0.92 \\ 0.24 \\ -0.79 \end{bmatrix} P = \begin{bmatrix} 7.7 & -1.26 & 6.18 & -1.94 & 7.94 & -4.37 & 1.72 \\ 0 & -6.19 & 2.5 & -6.24 & 0.73 & -4.41 & 8.29 \\ 0 & 0 & -3.89 & 2.35 & -4.02 & -0.91 & 2.76 \end{bmatrix}^T \quad (3.2)$$

Once the state of the economy, the factors and the macroeconomic variables have been simulated, the realizations of the macroeconomic variables were used to estimate the 3 factors using the Kalman Filter. The results for the simulations and the estimated factors are presented in Figure 3.9 and Figure 3.10.

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 simulated. 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.

After running the genetic algorithms subroutine, the optimal value for this case was  $z^* = 1.707$ , corresponding to a 0-coincidence p-value of 1 and a correlation of 0.707. This optimal value is attained at  $\theta^* = 2.8625, \varphi^* = 1.6562$  in spherical coordinates, and

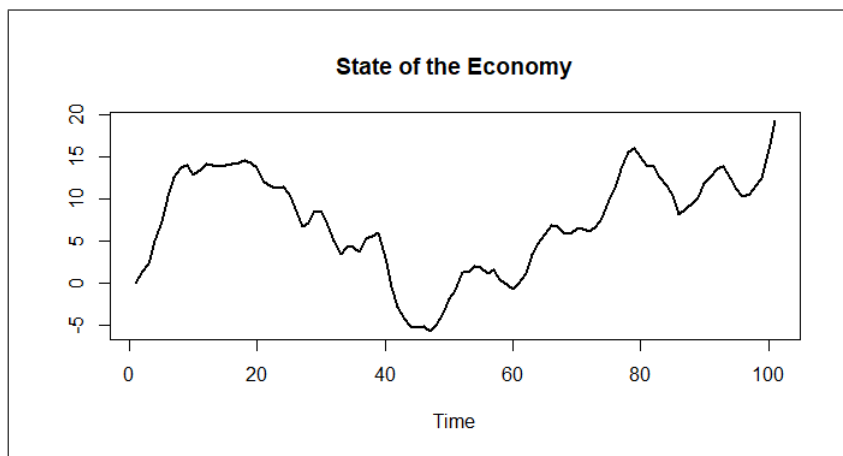


FIGURE 3.9. Simulation of the state of the economy (3D).

$\alpha_1^* = -0.957, \alpha_2^* = -0.277, \alpha_3^* = -0.085$  in Cartesian coordinates. Figure 3.11 shows the objective function over the whole spectrum and Figure 3.12 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.

Based on the resampling technique proposed, a 95% unilateral confidence interval for the optimal value of the objective function is  $[1.097, 2]$ . Figure 3.13 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  $(\theta, \varphi) = (\pi/2, \pi/2)$  belongs to the 95% confidence region. In fact, the performance measures for each one of the factors are presented in Table 3.2.

TABLE 3.2. Performance measures for the single factors (3D)

| Performance measure | Factor 1 | Factor 2 | Factor 3 |
|---------------------|----------|----------|----------|
| $z^*$               | 0.052    | 1.143    | 0.54     |
| <i>p-value</i>      | 0.75     | 1        | 0.5      |
| <i>cor</i>          | -0.698   | 0.143    | 0.04     |

### 3.3 Multiple Replications of the Simulation Framework

The procedure presented in the previous two sections was replicated 100 times starting from random and independent base case scenarios for both the two-dimensional and the three-dimensional framework.

For each one of the replications, the following information was recorded:

- 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,

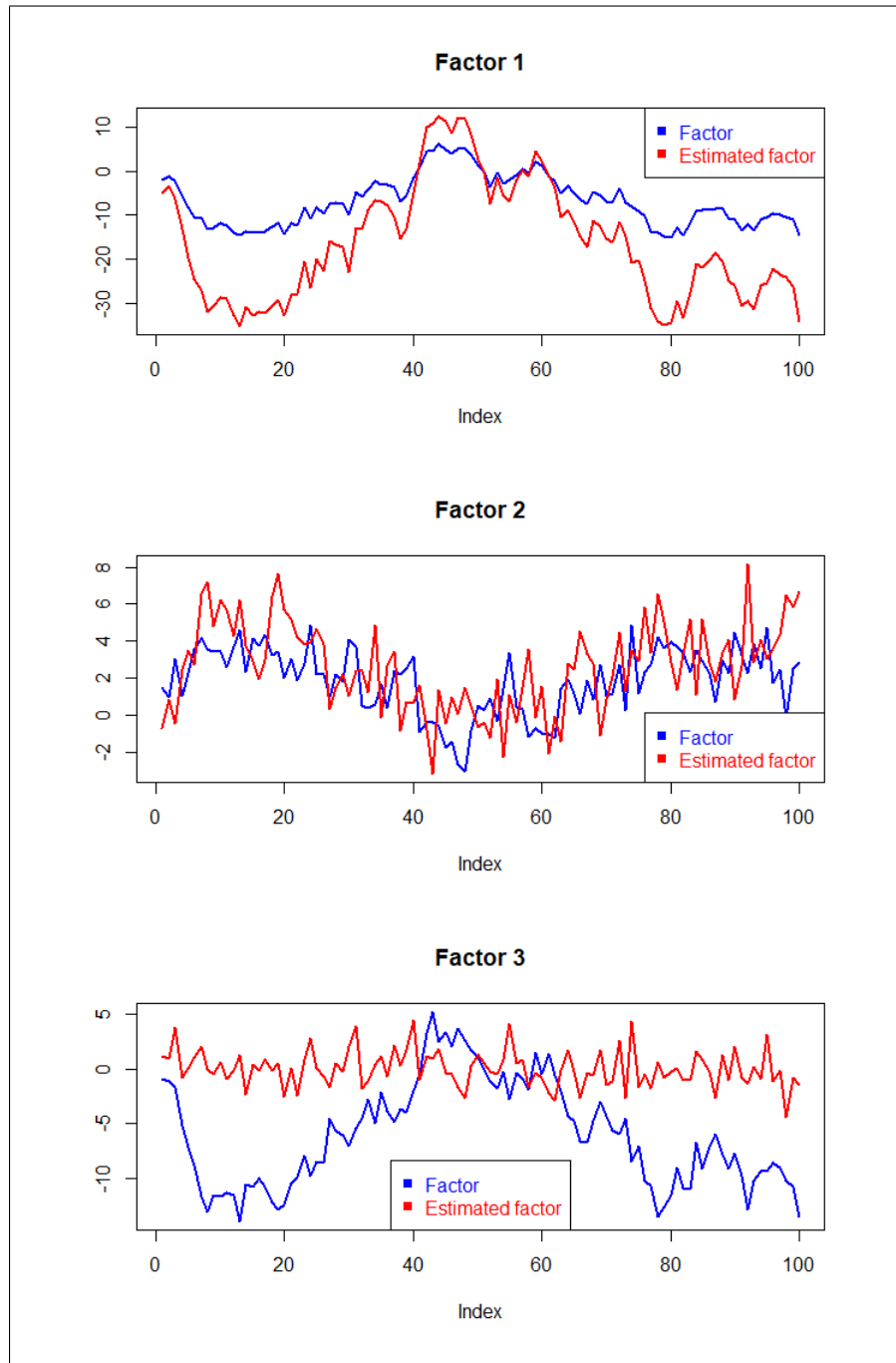


FIGURE 3.10. Comparison of the simulated versus the estimated factors (3D).

- 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

$$gap := z^* - \max_{l \in \{1, 2, \dots, r\}} p(\Delta \hat{c}_t, \Delta \hat{f}_{lt}), \quad (3.3)$$

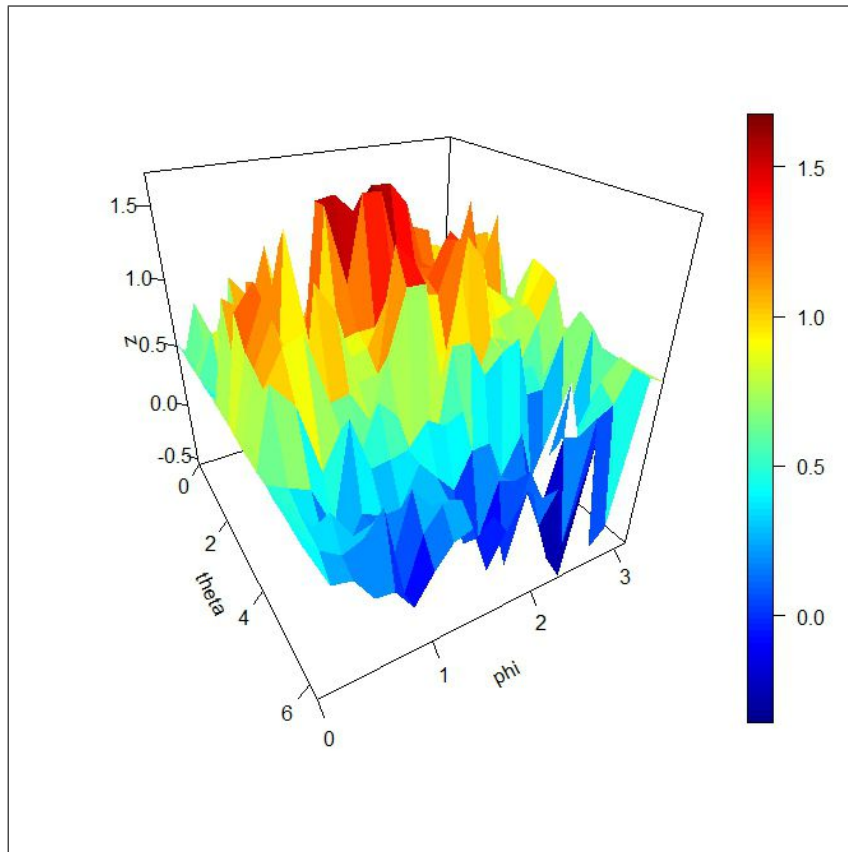


FIGURE 3.11. Representation of the objective function in spherical coordinates (3D).

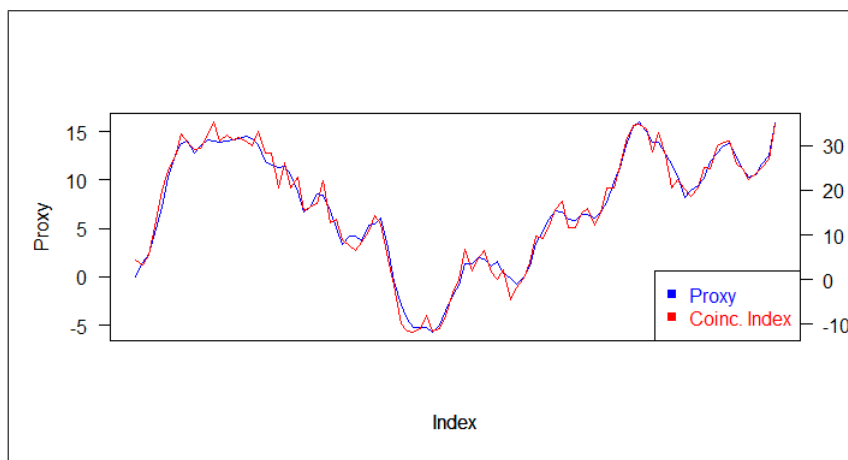


FIGURE 3.12. Comparison between the realization of the state of the economy (proxy) and the coincident index (3D).

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

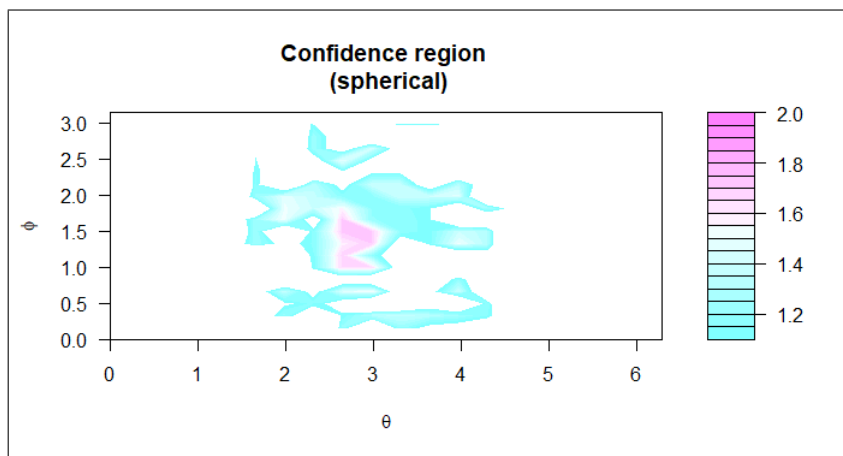


FIGURE 3.13. Confidence region in spherical coordinates (3D).

$$RI := \frac{gap}{\max_{l \in \{1, 2, \dots, r\}} p(\Delta \hat{c}_t, \Delta \hat{f}_{lt})}. \quad (3.4)$$

These quantities were then aggregated along 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. The results can be seen in table 3.3.

TABLE 3.3. Aggregate performance summary for the replications

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

For instance, for the 2D framework, 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 simultaneous cross-correlation) 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 supports the idea that one of the single factors itself can play the role of the coincident index.

In an analogous way, for the 3D framework, 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 importance of 42% approx.).

### 3.4 Application to Real Data

Nieto & Chudt (2017) computed a new coincident index for the Colombian economy based on the following six macroeconomic variables: (1) Industrial Production, (2) Electric Energy Consumption, (3) Production of Sugar Cane, (4) Retailing Commerce Excluding Fuels and Vehicles, (5) Cement Production, and (6) Unemployment Rate, during the period starting in January, 2000 until June, 2017.

They identified two common factors 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.

For this application, the same estimates of the factors were used in order to compare the results obtained. Figure 3.14 and Figure 3.15 show the proxy for the state of the economy and the two estimated factors.

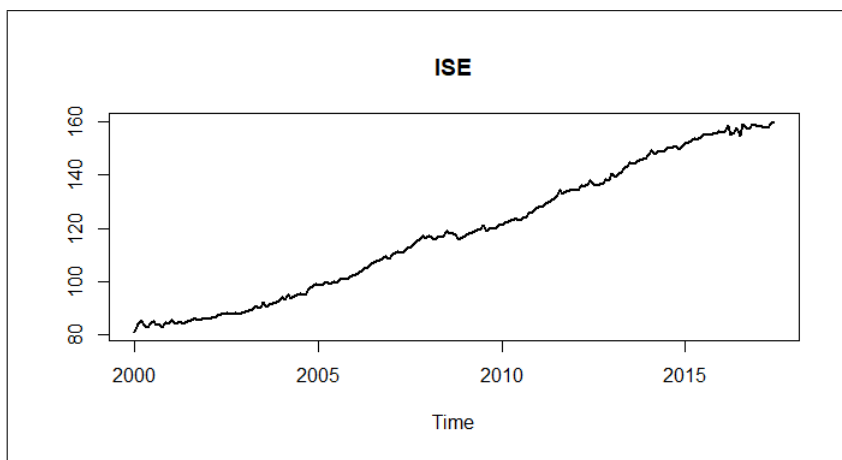


FIGURE 3.14. Computation of the ISE from Jan-2000 to June-2017.

The optimization process to estimate the coefficients of the linear combination 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  $\theta^* = 6.282$  in polar coordinates, and  $\alpha_1^* = 0.999$ ,  $\alpha_2^* = -0.002$  in Cartesian coordinates. The progress of the optimization procedure is shown in Figure 3.16, while the comparison between the proxy and the coincident index is presented in Figure 3.17.

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 this result, the 95%-confidence region for the coefficients of the linear combination is presented in Figure 3.18 and Figure 3.19. Based on this confidence region, it can be concluded that with a confidence level of 95% factor 1, with coefficients  $(\alpha_1, \alpha_2) = (1, 0)$ , can play the role of the coincident index. The performance measures for each of the individual factors are presented in Table 3.4.

The results for the application in the Colombian context are consistent to what had been previously obtained by Nieto & Chudt (2017) because both procedures have sug-

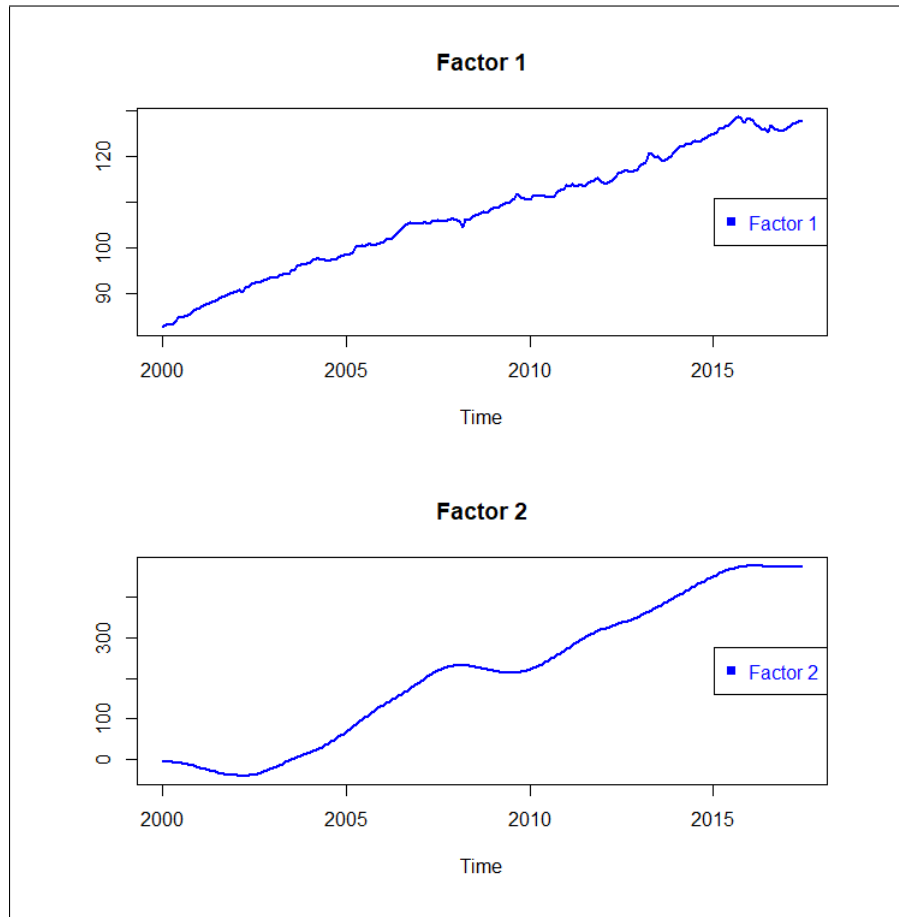


FIGURE 3.15. Factors estimated by Nieto et al (2017).

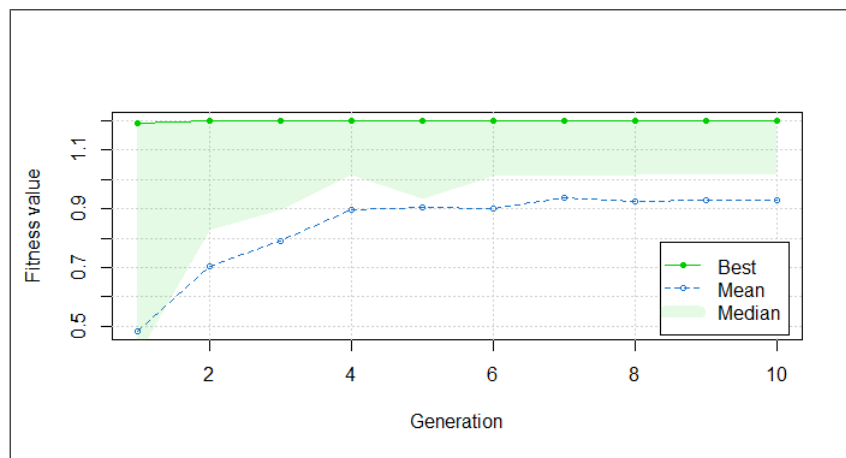


FIGURE 3.16. Progress of the optimization procedure across the generations (Real data).

gested that factor 1 can play the role of the coincident index for the Colombian economy. Nevertheless, the primary value that the new methodology presented here offers is the possibility to expand the set of candidates to be coincident index and actually conclude,

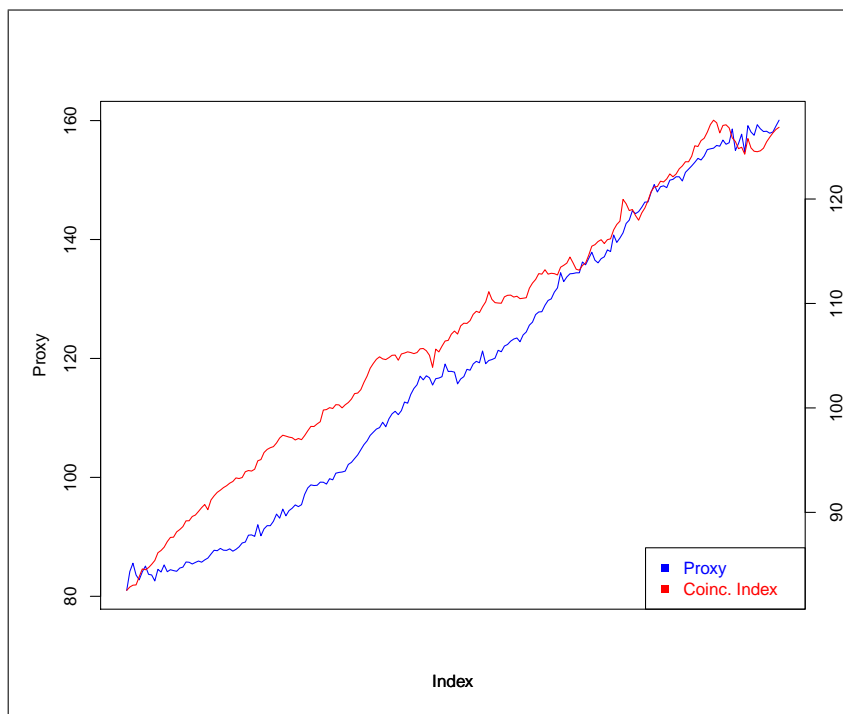


FIGURE 3.17. Comparison between the proxy and estimated coincident index (Real data).

TABLE 3.4. Performance measures for the single factors (Real data)

| Performance measure | Factor 1 | Factor 2 |
|---------------------|----------|----------|
| $z^*$               | 1.2      | 0.771    |
| $p$ -value          | 0.875    | 0.633    |
| $cor$               | 0.325    | 0.138    |

via a statistical test with a pre-specified significance level that, among all the possible linear combinations of factors, factor 1 can be considered as a coincident index.

In fact, this approach can be considered as a generalization of what Martínez et al. (2016) and Nieto & Chudt (2017) have published when only the 0-coincidence is of interest to select a coincident index.

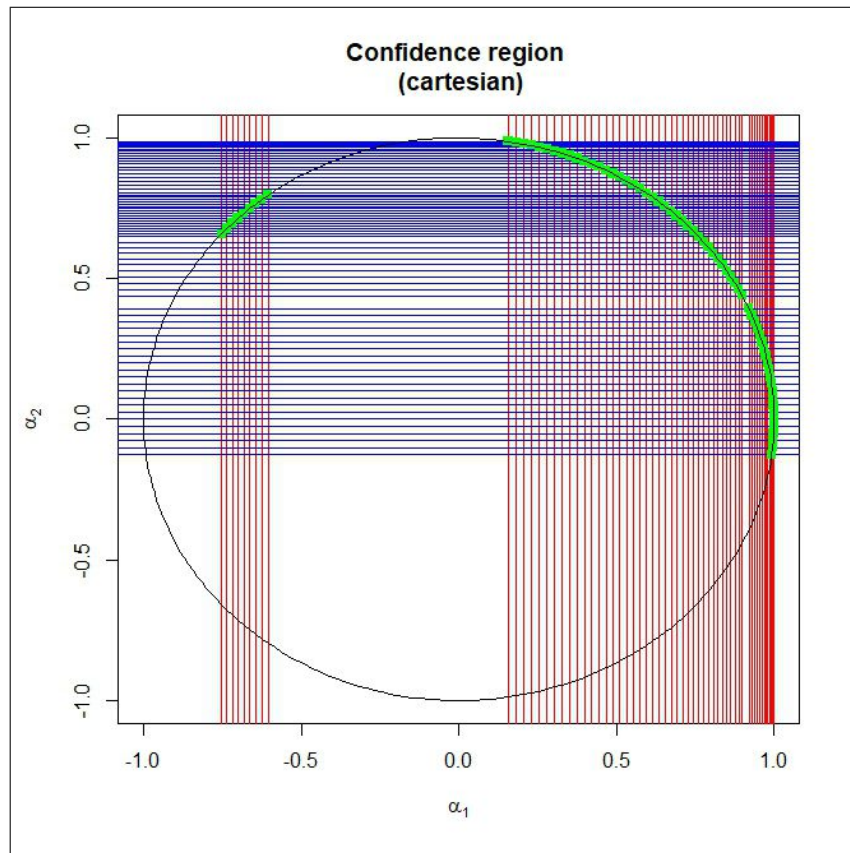


FIGURE 3.18. Confidence region in Cartesian coordinates (Real data).

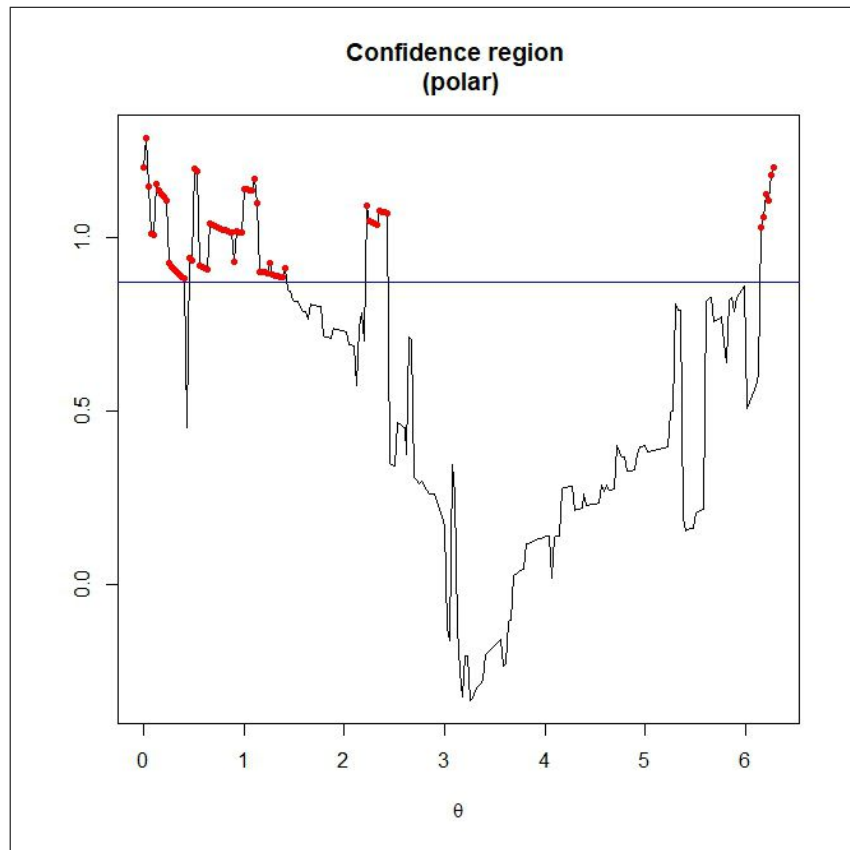


FIGURE 3.19. Confidence region in polar coordinates (Real data).

---

---

## Conclusions and Recommendations

---

---

This work proposes a novel methodology to design coincident indices as linear combinations of the common dynamic factors from a multivariate time series of macroeconomic variables. The procedure presented is optimal in the sense of maximizing the sum of two statistical indicators of goodness of performance in the context of macroeconomic indices: the 0-coincidence p-value and the cross-correlation at lag 0.

Besides, this methodology offers and justifies a set of statistical inference procedures to conduct point estimates, interval estimates and hypothesis tests to assess if any linear combination of factors can behave as a coincident index for the state of the economy.

Even if this new methodology is presented as an improvement of one of the phases in Martínez et al's methodology (Martínez et al., 2016), it does not have to be applied only under Martínez's assumptions since the linear combination construction starts by taking the estimated factors as input. The entire statistical framework can be easily adapted and linked to other methodologies to estimate common factors in multivariate time series, such as Stock & Watson (2011), Forni et al. (2005), Lam et al. (2012), Bujosa et al. (2013), among others.

The great advantage of using the factors from Martínez et al's methodology is the fact that they account for the potential nonstationarity condition usually exhibited by the time series (Martínez et al., 2016). Despite the methodology presented in this work considers first differences of the proxy and the index in the objective to be maximized, the index itself is defined as a function of both nonstationary and stationary common factors coming from Martínez et al's methodology in an attempt to account for the nonstationarity of the original series. Nevertheless, the way in which the nonstationarity is taken into account needs to be further studied.

On the other hand, the simulated scenarios cast 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.
- (2) The estimated factors usually come with opposite signs, requiring the use of negative constants to actually mimic the behavior of the proxy.

Additionally, the methodology was applied to the Colombian context producing consistent results to what had been previously obtained by some of the researchers consulted,

---

and offering statistical validation to previous conclusions by means of the analysis of the simulated sampling distribution for the linear-combination estimators.

Finally, a complete implementation of the procedures involved was developed in R<sup>®</sup> and the source codes are available to anyone interested upon request to the author via e-mail.

However, there are other aspects that are worth revising in future research efforts. The following list presents, in a non-exhaustive way, some of the potential new directions that can be further studied to improve this methodology:

- Analyze the algorithmic structure of the methodology from the stand point of computer efficiency.
- Include some other steps that have to be actually considered in a real application, such as the identification of the number and type of factors, the identification of their evolution dynamics, i.e., their VARIMA structure, their estimation framework, among others; to examine their potential impacts on the overall outcome.
- Consider weighted averages of the two elements in the objective function or different components in the objective function. For instance other measures of  $i$ -coincidence,  $i \neq 0$ .
- Consider non-linear combinations of the common factors to create the economic index.

---

---

## Bibliography

---

---

- Albelwi, S. & Mahmood, A. (2017). A framework for designing the architectures of deep convolutional neural networks, *Entropy* **19**(6): 242.
- Altissimo, F., Cristadoro, R., Forni, M., Lippi, M. & Veronese, G. (2010). New euro-coin: tracking economic growth in real time, *The review of economics and statistics* **92**(4): 1024–1034.
- Banerji, A. (1999). The lead profile and other non-parametric tools to evaluate survey series as leading indicators, *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.
- Bethke, A. D. (1978). *Genetic algorithms as function optimizers*, PhD thesis.
- Bickel, P. J. & Doksum, K. A. (2015). *Mathematical statistics: basic ideas and selected topics, volume I*, Vol. 117, CRC Press.
- Bloch, E. D. (2011). *Proofs and fundamentals: a first course in abstract mathematics*, Springer Science & Business Media.
- 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. et al. (1946). Measuring business cycles, *Nber Books* .
- Cristadoro, R., Forni, M., Reichlin, L. & Veronese, G. (2005). A core inflation indicator for the euro area, *Journal of Money, credit, and Banking* **37**(3): 539–560.
- Escribano, A. & Peña, D. (1994). Cointegration and common factors, *Journal of Time Series Analysis* **15**(6): 577–586.
- Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation, *Review of Economics and statistics* **82**(4): 540–554.

- 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).
- Johnson, R. A. & Wichern, D. (2002). *Multivariate analysis*, Wiley Online Library.
- 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.
- Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization, *The computer journal* **7**(4): 308–313.
- Nieto, F. H. & Chudt, N. (2017). Construcción de un índice coincidente para la actividad económica colombiana.
- 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.
- 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.
- Sastry, K., Goldberg, D. E. & Kendall, G. (2014). Genetic algorithms, *Search methodologies*, Springer, pp. 93–117.
- 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. (2011). Dynamic factor models, *Oxford handbook on economic forecasting* .
- Stock, J. H. & Watson, M. W. (1988). A probability model of the coincident economic indicators.
- Stock, J. H. & Watson, M. W. (1989). New indexes of coincident and leading economic indicators, *NBER macroeconomics annual* **4**: 351–394.

---

Stock, J. H. & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors, *Journal of the American statistical association* **97**(460): 1167–1179.

*Using Genetic Programming to evolve Trading Strategies* (n.d).  
<http://www.turingfinance.com/using-genetic-programming-to-evolve-security-analysis-decision-trees/>.

**URL:** <http://www.turingfinance.com/using-genetic-programming-to-evolve-security-analysis-decision-trees/>

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