UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

TEORÍA DE MATRICES ALEATORIAS APLICADA AL ANÁLISIS ESTADÍSTICO DE UN MODELO DE FACTORES

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN
CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

CAMILA FERNANDA BRITO PIZARRO

PROFESOR GUÍA:
DANIEL REMENIK ZISIS
JOAQUÍN FONTBONA TORRES

MIEMBROS DE LA COMISIÓN:
JAIME SAN MARTÍN ARISTEGUI
GUSTAVO SOTO MUSTER

Este trabajo ha sido parcialmente financiado por Núcleo Milenio:
"MODELOS ESTOCÁSTICOS DE SISTEMAS COMPLEJOS Y DESORDENADOS".

SANTIAGO DE CHILE
2016

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE: INGENIERO CIVIL MATEMÁTICO
POR: CAMILA FERNANDA BRITO PIZARRO
FECHA: 18 de julio de 2016
PROF. GUÍA: DANIEL REMENIK ZISIS
PROF. GUÍA: JOAQUÍN FONTBONA TORRES

TEORÍA DE MATRICES ALEATORIAS APLICADA AL ANÁLISIS ESTADÍSTICO DE UN MODELO DE FACTORES

Las matrices aleatorias y su reciente teoría están jugando un papel fundamental como herramienta estadística en áreas tales como finanzas, meteorología y procesamiento de señales e imágenes. Algunas de las aplicaciones que han adquirido mayor desarrollo se encuentran en el sector financiero y en el área de las comunicaciones inalámbricas. El desafío planteado en este trabajo de tesis consiste en realizar un análisis estadístico basado en la teoría de matrices aleatorias referido a un modelo de factores. A través de la experimentación computacional, se pretende alcanzar dos metas.

La primera de ellas consiste en contrastar dos versiones de un mismo test de hipótesis, las cuales se definen a partir de estadísticos provenientes de dos de las más conocidas familias gaussianas de matrices aleatorias: GUE y GOE. Esta comparación surge del hecho de que la familia GOE es menos estudiada en las aplicaciones de matrices aleatorias a considerar, de modo que se busca ampliar el conocimiento que de ella se tiene. Para hacer efectivo el contraste entre ambas versiones, estas se implementan para luego analizarlas en términos de sus comportamientos frente a errores y aciertos. Así, se logra probar empíricamente que no existe diferencia alguna entre ellas, por lo que la versión GOE del test es la que asume el protagonismo.

Alcanzada la meta anterior, la segunda consiste en dar utilidad al test en su versión GOE, mediante el desarrollo de un procedimiento que lo aplica iteradas veces para estimar el número de factores de una muestra sujeta al modelo de factores. Posteriormente, el procedimiento es sometido a una serie de pruebas empíricas que buscan validarlo como método de estimación del número de factores.

Finalmente, es preciso mencionar que, si bien, este trabajo posee un carácter fundamentalmente experimental, no se aparta del estudio, análisis y manejo abstracto de la teoría de matrices aleatorias que se requieren necesariamente para llevarlo a cabo.

A Dios y a mi familia.

Agradecimientos

A mi mamá por amarme y cuidarme de la forma en que solo ella lo hace. Por darme fuerzas y estar siempre al pie del cañón conmigo sobre todo en mis momentos más difíciles. A mi papá, por siempre entenderme y confiar en mí, por ser siempre mi cable a tierra y el que me aconseja bien. Les agradezco a ambos porque me lo han dado todo a mi y a mi hermano, agradezco la confianza, el amor y el apoyo que me dieron para llegar hasta acá. Me siento agradecida de que sean mis papás y de que estén conmigo. Esto es más de uds que mío.

A Cristóbal, mi perrito, gracias por quererme y devolverme a la niñez cada vez que estoy contigo.

A Andrés, mi mejor compañero, gracias por ser como eres, por enseñarme cómo ser feliz, por aguantarme y entenderme aún cuando ni yo me entiendo. Le agradezco a Dios por ponerte en mi vida.

A Eduardo, mi gran amigo, gracias por pedirme ese cuaderno de proba porque me permitió conocer a un gran partner, el mejor de todos. Gracias por nuestras largas charlas, nuestras peleas incansables por quién sufre más, nuestras aventuras en las prácticas y por ayudarme siempre de corazón en todo. Sin duda no hubiese sido lo mismo sin ti. Somos un equipazo.

A mis amigos de la carrera, Waldo por ayudarme siempre con todo cada vez que le hacía preguntas, a Hugo por siempre sacarme una risa, a Giancarlo por su humor, Juan por su buena onda y Edgardo por ayudarme hasta con mis dudas existenciales.

A mis amigos de la sección, Feña por ser mi confidente todos estos años, Patito por ser tan chevere como eres, Alex por ser siempre el más apañador, Herni y Marcos por transmitir siempre la buena onda, Toño por ser un amigo bkn, Nacho por esperarme para almorzar después de las largas clases de análisis y a Rodrigo por escucharnos nuestras penas y alegrías siempre.

A los amigos de ahora, Caro, Pichón, Fran, Rodrigo y Gonzalo por querer como quieren a nuestro Erick y por integrarme también.

A la señora Elsa y a la señora Iris porque ambas fueron un gran apoyo en mis inicios en la U. Siempre recordaré lo que hicieron por mi y por mi familia.

A la comisión por tener la mejor disposición conmigo, en especial a los profesores Daniel y Joaquín por darme la oportunidad de trabajar con ellos y de integrarme al núcleo.

Tabla de Contenido

Índice de tablas

Índice de Ilustraciones

Introducción

0.1. Estado del Arte

Hoy por hoy, las ciencias atmosféricas, el tratamiento de imágenes médicas, el área de las comunicaciones y la economía, por nombrar sólo algunas, se ven enfrentadas al problema de lidiar con el manejo de grandes volúmenes de datos, lo que se denomina *Análisis de datos de gran dimensión* (o *LDDA*, por sus siglas en inglés). En este contexto, la estadística juega un papel importante, pues debe responder a la creciente necesidad de trabajar con datos de gran dimesión.

Es más, la estadística hoy se encuentra en una etapa en que no sólo debe manejar *LDDA*, sino que va más allá, pues, tiene como objetivo estudiar la validez de los teoremas límites clásicos de estadística multivariada, en aquellos casos en que se quiere observar el comportamiento asintótico no sólo del tamaño de la muestra, sino también de la dimensión de la matriz de datos. En este escenario surge la teoría de matrices aleatorias (o *RMT* por *Random Matrix Theory*) cuyas aplicaciones estadísticas van desde herramientas y métodos de reducción de dimensiones hasta *clustering* [7] y test de hipótesis [15].

Se ha observado que el comportamiento límite de la dimensión en matrices aleatorias aparece en las áreas antes descritas, como leyes que emergen del modelamiento matemático dado. En particular, es de especial interés el análisis espectral límite de las matrices aleatorias, puesto que varios estadísticos tienen estrecha relación con la distribución del espectro.

En este orden de ideas, es pertinente mencionar que en la década de los 80 se han efectuado contribuciones respecto del trabajo de [2] y [11], en relación a la existencia de una distribución espectral límite de matrices aleatorias y la forma explícita que esta tiene para cierta clase de matrices. Recientemente, el interés se enfoca en estudiar teoremas límites de segundo orden, como es el teorema central del límite [5] y los concernientes a distribuciones límites del espectro y a valores propios extremos.

0.2. Propuesta

Para la construcción de la propuesta cobran relevancia, por tener estrecha relación, los siguientes antecedentes:

Una de las áreas en las que más literatura se puede encontrar sobre teoría de matrices aleatorias -en adelante *RMT*- es en macroeconomía y finanzas, particularmente en el contexto del modelo de factores.

Para los economistas, el principal objetivo siempre ha sido obtener la mayor información contenida en sus datos con el menor uso de recursos, y es por esto que, a menudo, se utilizan técnicas estadísticas como *PCA*, análisis factorial, entre otras, para conservar precisamente la mayor cantidad de información dentro de unos pocos *factores*. En consecuencia, se hace necesario conocer el número de factores que rigen los datos, pues muchas veces este tiene interesantes interpretaciones económicas y se requiere para realizar estimaciones o pronósticos.

En consideración a lo anterior, la propuesta que contiene y se presenta en este trabajo de tesis, corresponde al desarrollo e implementación de un procedimiento que permita estimar el número de factores de una muestra. Este procedimiento se basa en la aplicación iterativa de un test de hipótesis, cuyo estadístico, visto como variable aleatoria, tiene estrecha relación con el comportamiento del espectro de matrices pertenecientes a la familia *GOE*. Por lo demás, tal procedimiento constituye una innovación, pues solo existe una versión que utiliza un test basado en el espectro de la familia *GUE* que es más simple de estudiar teóricamente.

En la práctica, esta labor se traduce en que, al estudiar las leyes del espectro de estos *ensembles*, es posible comprender las leyes de los estadísticos asociados, y por ende, su funcionamiento. Luego, este conocimiento teórico da paso a la experimentación que pondrá a prueba al procedimiento y al test sobre el cual se soporta.

Para llevar a cabo la citada propuesta, el desarrollo de este trabajo comprende tres capítulos. El primero contiene los antecedentes teóricos que atañen a teoría de matrices aleatorias necesarios para el caso. El segundo se encarga de detallar el modelo de factores y los supuestos con los que se trabaja, de describir las versiones *GUE* y *GOE* del test de hipótesis, y de introducir el procedimiento de estimación del número de factores. Por último, el tercer capítulo se refiere a la implementación y al estudio del test *GOE* y del procedimiento, a las pruebas empíricas a las que son sometidos con sus respectivos resultados y a la descripción de una aplicación del procedimiento a datos reales.

Capítulo 1

Teoría básica de Matrices Aleatorias

En esta sección se detallan las definiciones y la teoría básica con la que se trabajará en la tesis. Se trata, en otras palabras, de un breve resumen teórico atinente a matrices aleatorias.

1.1. Familias de Matrices Aleatorias

Las matrices aleatorias fueron introducidas alrededor del año 1950 por E. Wigner en un contexto de física nuclear, en el cual los niveles de energía de átomos pesados son descritos a través de un operador Hamiltoniano que E.Wigner modeló como una matriz aleatoria.

En este trabajo de tesis se destacan dos grandes familias clásicas de matrices aleatorias.

Definición 1.1 (**Familias clásicas**) *Se definen a continuación dos de las más conocidas familias de matrices aleatorias.*

1. *Una matriz H de dimensiones $n \times n$ pertenece a la clase de matrices aleatorias $\beta = 1$ si:*

   (i) Es real simétrica, es decir, $H = H^t$.

   (ii) Tiene $\frac{n(n+1)}{2}$ entradas independientes (las restantes entradas vienen dadas por simetría).

   (iii) H es diagonalizable por una transformación ortogonal.

2. *Una matriz H de dimensiones $n \times n$ pertenece a la clase de matrices aleatorias $\beta = 2$ si:*

   (i) Es hermitiana compleja, es decir, $H_{ij} = \overline{H}_{ji}$ para $i, j = 1, \ldots, n$ y $H_{ij} \in \mathbb{C}$. Lo que se denota como $H = H^*$.

   (ii) Tiene n^2 entradas independientes.

(iii) Es diagonalizable por una transformación unitaria.

Definición 1.2 (**Familias Gaussianas**) *Se definen, además, las familias de matrices aleatorias gaussianas siguientes:*

1. ***Familia Gaussiana Ortogonal:*** *(Gaussian Ortogonal Ensemble -en adelante GOE-) Corresponde a la familia de matrices aleatorias H de $n \times n$ que son reales simétricas con medida de probabilidad*

$$p(H)dH = \frac{1}{Z_{GOE(n)}} \mathrm{e}^{-\frac{n}{4} tr(H^2)} dH \quad ,$$

con $Z_{GOE(n)}$ una constante de normalización y $dH = \prod_{i=1}^{n} dH_{ii} \prod_{1 \leq i < j \leq n} dH_{ij}$ el análogo de la medida Lebesgue sobre el espacio de las matrices reales simétricas.
A esta familia (o ensemble) se le asocia el índice $\beta = 1$.

2. ***Familia Gaussiana Unitaria:*** *(Gaussian Unitary Ensemble -GUE en adelante-) Corresponde a la familia de matrices H complejas de $n \times n$ que son hermitianas y con medida de probabilidad*

$$p(H)dH = \frac{1}{Z_{GUE(n)}} \mathrm{e}^{\frac{-n}{2} tr(H^2)} dH$$

con $Z_{GUE(n)}$ una constante de normalización y $dH = \prod_{i=1}^{n} dH_{ii} \prod_{1 \leq i < j \leq n} d\mathrm{Re}(H_{ij}) d\mathrm{Im}(H_{ij})$ el análogo a la medida de Lebesgue en el espacio de las matrices hermitianas.
A esta familia (o ensemble) se le asocia el índice $\beta = 2$.

Observación: Existe, además, una familia asociada al índice $\beta = 4$ llamada Gaussiana Simpléctica *(Gaussian Symplectic Ensemble)* o *GSE* abreviadamente que corresponde al espacio de las matrices de cuaterniones de $n \times n$, cuya definición no será detallada en este trabajo de tesis por no ser necesaria.

Es común ver en la literatura de *RMT* una especie de dicotomía entre las familias *GUE* y *GOE* en relación a las pruebas teóricas de sus propiedades. Por un lado, las características y resultados más clásicos sobre la familia *GUE* han sido rápidamente estudiados y probados, y por el otro, no ocurre lo mismo con la familia *GOE*. Si bien este último caso es también estudiado, es difícil dar demostraciones a sus versiones de los resultados en cuestión, por lo que estas aparecen en la literatura tardíamente, debido a que el estudio analítico de las familias *GOE* posee una dificultad matemática mayor que el del *GUE*.

Otras matrices notables que aparecen frecuentemente en el estudio de la *RMT* son las denominadas matrices *Wigner* y *Wishart* que se definen a continuación:

Definición 1.3 (**Matriz de Wigner**) *Considérese una familia de v.a.* $\{Z_{i,j}\}_{1\leq i<j}$ *independientes, de media cero, a valores reales o complejos. Considérese además, independiente de* $Z_{i,j}$*, otra familia* $\{Y_i\}_{1\leq i}$ *de v.a. i.i.d, de media cero y a valores reales. Una matriz de Wigner se define como la matriz* X_n *de tamaño* $n \times n$ *cuyas entradas son:*

$$X_n(j,i) = \overline{X}_n(i,j) = \begin{cases} Z_{i,j} & si\ i < j \\ Y_i & si\ i = j \end{cases}$$

en donde $\overline{X}_n$ *denota el conjugado de* X_n*.*

Observación: Si las v.a. $Z_{i,j}$ e Y_i tienen distribución gaussiana, entonces se dice que X_n es una *matriz de Wigner gaussiana.*

Definición 1.4 (**Matriz Wishart**) *Una matriz Wishart* W *con covarianza* Σ *y* n *grados de libertad se define como:*

$$W = \sum_{i=1}^{n} Z_i Z_i',$$

donde Z_i *se distribuye i.i.d* $\mathcal{N}(0,\Sigma)$*.*

Observación: De forma matricial si Z es una matriz de $n \times p$ con entradas i.i.d $\mathcal{N}(0,1)$ y Σ es de dimensiones $p \times p$, entonces $W = \Sigma^{1/2} Z' Z^{1/2} \sim \mathcal{W}_p(\Sigma, n)$, equivalentemente se dice que W tiene una distribución Wishart con n grados de libertad.

Ejemplo 1 [Aplicado al problema de detección de un target]
La detección de un objetivo (o *target*) inmerso en un medio, es un problema tipo en el tratamiento de imágenes, especialmente en aquellos casos en que se utilizan sensores de onda. Para estudiarlo, [10] propone insertar en el medio una serie de receptores y fuentes emisoras de ondas. Las respuestas entre cada par de receptor y fuente constituyen una *matriz de respuesta*. El inconveniente surge cuando los datos están alterados debido a ruido que se asume típicamente aditivo. El gran objetivo es seguir detectando el *target* en la presencia de ruido, y es aquí donde entra toda la maquinaria de la *RMT*.

Considérense M fuentes y N receptores, y denótese como A_0 a la matriz de respuesta de dimensiones $N \times M$. Por otra parte, debido a la presencia de ruido, la señal medida por cada receptor está corrompida por un ruido aditivo que se modela en términos de una v.a. compleja que se asume $\mathcal{N}(0,\sigma^2)$. Matemáticamente, esto quiere decir que la parte real y la parte imaginaria del ruido son independientes y siguen una distribución $\mathcal{N}(0,\frac{\sigma^2}{2})$.

Se realizan M experimentos en los que en el m-ésimo de ellos ($m = 1, 2, \ldots, M$), la m-ésima fuente emite una señal de onda que es captada por los N receptores. La información que se obtiene puede codificarse en una matriz A que corresponde a la suma de la matriz de respuesta no perturbada (es decir, sin ruido) y el ruido. En otras palabras, lo que se tiene es

$$A_{nm} = A_{0,nm} + W_{nm} \quad n = 1, ..., N, \quad m = 1, ..., M, \tag{1.1}$$

donde W_{nm} son v.a. complejas i.i.d. con distribución $\mathcal{N}(0,\sigma^2)$ que representan el ruido.

En este contexto, al definir la matriz $S = \frac{1}{M} WW^T$, se desprende que corresponde a una matriz de Wishart normalizada.

1.2. Universalidad y estudio del espectro

Durante el desarrollo de la *RMT* surge un concepto que en la literatura se conoce como *universalidad*. Esta noción alude a un fenómeno que se presenta repetidamente en la naturaleza. Así, por ejemplo, se ha observado que los tiempos de llegada de buses a un paradero en la ciudad mexicana de Cuernavaca [14] y la dispersión de los neutrones en un núcleo atómico como el Uranio238, entendidos como un proceso puntual, tienden a distribuirse y repelerse entre sí tal como lo hacen los valores propios de una matriz aleatoria. Esta universalidad, entre otras cosas, es la principal motivación para hacer del espectro de las matrices aleatorias un tema de estudio a cabalidad.

Una rama importante de la *RMT* se dedica al estudio asintótico de la teoría espectral, es decir, a estudiar la distribución de los valores propios cuando la dimensión de la matriz tiende a infinito.

Definición 1.5 (**Medida Espectral Empírica**) *Sea H_n matriz de dimensiones $n \times n$, real simétrica o hermitiana, y sean $\lambda_1^{(n)}, \dots \lambda_n^{(n)}$ sus valores propios (donde el superíndice (n) alude a la dimensión de la matriz).*

Se define la medida espectral empírica como:

$$\mu_{H_n}(A) = \frac{1}{n} \left| \{ j \in \{1, \dots, n\} : \lambda_j^{(n)} \in A \} \right| \quad A \subset \mathbb{R}$$

Esta medida espectral tiene una ley límite para matrices aleatorias Wigner y Wishart, lo que se conoce clásicamente como *Ley del semicírculo de Wigner* y *Ley de Marchenko-Pastur* respectivamente.

Teorema 1.6 (**Ley del Semicírculo de Wigner**) *Sea X una matriz aleatoria de Wigner de dimensiones $n \times n$ y cuyas entradas tienen distribución $\mathcal{N}(0,1)$. La medida espectral empírica esperada de la matriz $\frac{1}{\sqrt{n}} X$ tiende, en el regimen $n \to \infty$, a la denominada ley del semicírculo, denotada como F y cuya densidad viene dada por*

$$dF(x) = \frac{1}{2\pi} \sqrt{4 - x^2} \mathbb{1}_{[-2,2]}(x).$$

Observación: (i) Para el caso en que la distribución de las entradas de la matriz de Wigner no sea estándar, es decir, para el caso en que la distribución sea $\mathcal{N}(0, \sigma)$, la ley del semicírculo sigue siendo válida con una densidad dada por

$$dF(x) = \frac{1}{2\pi\sigma^2} \sqrt{4\sigma^2 - x^2} \mathbb{1}_{[-2\sigma, 2\sigma]}(x), \tag{1.2}$$

en donde $[-2\sigma, 2\sigma]$ corresponde al soporte en el que se distribuyen todos los valores propios.

(ii) Este resultado ha sido probado usando varias técnicas como la transformada de Stieljes y el método de los momentos (en [3] se resumen brevemente dichas demostraciones). Además se ha extendido en [2] y [11] para convergencia casi segura y en probabilidad bajo las condiciones apropiadas.

Ahora, en el contexto de una matriz de Wishart, en donde la dimensión (p) tiende a infinito proporcionalmente con los grados de libertad (n), es decir, se cumple $\frac{p}{n} = \gamma \in (0, \infty)$, se tiene el siguiente resultado descrito en [3] sobre la convergencia de la medida espectral empírica:

Teorema 1.7 (**Ley Marchenko-Pastur**) *La medida espectral empírica converge en probabilidad a la distribución denominada ley Marchenko-Pastur, cuya densidad viene dada por:*

$$p_\gamma(x) = \frac{1}{2\gamma\pi x\sigma^2}\sqrt{(b-x)(x-a)}\mathbb{1}_{[a,b]}(x) \ ,$$

en donde $a = \sigma^2(1-\sqrt{\gamma})^2$, $b = \sigma^2(1+\sqrt{\gamma})^2$.

Observación: (i) σ corresponde a la varianza de la distribución suyacente y en el caso en que $\sigma^2 = 1$ se habla de *Ley Marchenko-Pastur Estándar*.

(ii) Es posible obtener una versión más débil de este resultado como el que se prueba en [21].

Ejemplo 2 (**Aplicado al problema de detección de un target**) A modo de ejemplo se aplicará la ley de Marchenko-Pastur a uno de los resultados mostrados en [10] . Tal resultado se refiere al problema de detección de un objetivo que fue contextualizado en el Ejemplo 1 .

En el caso particular en que no existe un objetivo ($A_0 \equiv 0$ en el modelo 1.1) , en que solo existe ruido, es decir, la matriz en cuestión (A) consiste en coeficientes de ruido de media 0 y varianza $\frac{\sigma^2}{M}$, y en que el número de receptores supere la cantidad de fuentes ($N \geq M$), se denota como $\sigma_1^{(M)} \geq \sigma_2^{(M)} \geq ... \geq \sigma_M^{(M)}$ a los valores singulares de la matriz A, y como $\Lambda^{(M)}$ a la medida espectral. Formalmente,

$$\Lambda^{(M)}([a,b]) := \frac{1}{M}\left|\{j \leq M : \sigma_j^{(M)} \in [a,b]\}\right|$$

O, visto de otra forma como suma de masas de Dirac,

$$\Lambda^{(M)} = \frac{1}{M}\sum_{j=1}^{M}\delta_{\sigma_j^{(M)}}.$$

Entonces, el ejemplo consiste en probar la siguiente afirmación:

"La medida espectral $\Lambda^{(M)}$ *converge c.s. a la medida determinista* Λ, *definida por*

$$\Lambda([\sigma_u, \sigma_v]) = \int_{\sigma_u}^{\sigma_v}\frac{1}{\sigma}\rho_\gamma(\frac{s}{\sigma})ds, \tag{1.3}$$

donde ρ_γ es la denominada Ley del cuarto de círculo dada por

$$\rho_\gamma(x) = \frac{1}{\pi x\theta}\sqrt{((\sqrt{\gamma}+1)^2 - x^2)(x^2 - (\sqrt{\gamma}-1)^2)}\mathbb{1}_{(\gamma^{1/2}-1,\,\gamma^{1/2}+1]}(x)".$$

Demostración. Dado que los cuadrados de valores singulares de la matriz A corresponden a los valores propios de AA^*, tiene sentido calcular la ley de Marchenko-Pastur con el cambio de variable $u = x^2$, esto es,

$$\int p_\gamma(u)du$$

en donde p_γ es la densidad de la ley Marchenko-Pastur y se han omitido los límites de integración.

$$\begin{aligned}
\int p_\gamma(u)du &= \int 2xp_\gamma(x^2)dx \\
&= \int \cancel{2}x\frac{1}{\cancel{2}\pi x^2\gamma\sigma^2}\sqrt{(b-x^2)(x^2-a)}\mathbb{1}_{[a,b]}(x^2)dx \\
&= \int \frac{1}{\pi x\gamma\sigma^2}\sqrt{(b-x^2)(x^2-a)}\mathbb{1}_{[a,b]}(x^2)dx
\end{aligned}$$

Reemplazando los valores de a y b, se llega a

$$\begin{aligned}
&= \int \frac{1}{\pi\gamma x\sigma^2}\sqrt{\left(\sigma^2(\gamma^{1/2}+1)^2 - x^2\right)\left(x^2 - \sigma^2(\gamma^{1/2}-1)^2\right)}\mathbb{1}_{[\sigma^2(\gamma^{1/2}-1)^2,\sigma^2(\gamma^{1/2}+1)^2]}(x^2)dx \\
&= \int \frac{1}{\pi\gamma\cancel{\sigma^2}x}\sqrt{\cancel{\sigma^2}\left((\gamma^{1/2}+1)^2 - \frac{x^2}{\sigma^2}\right)\cancel{\sigma^2}\left(\frac{x^2}{\sigma^2} - (\gamma^{1/2}-1)^2\right)}\mathbb{1}_{[\sigma(\gamma^{1/2}-1),\sigma(\gamma^{1/2}+1)]}(x)dx \\
&= \int \frac{1}{\pi\gamma x}\sqrt{\left((\gamma^{1/2}+1)^2 - \frac{x^2}{\sigma^2}\right)\left(\frac{x^2}{\sigma^2} - (\gamma^{1/2}-1)^2\right)}\mathbb{1}_{[\gamma^{1/2}-1,\gamma^{1/2}+1]}(x/\sigma)dx \\
&= \int \underbrace{\frac{1}{\sigma}\frac{\sigma}{\pi\gamma x}\sqrt{\left((\gamma^{1/2}+1)^2 - \frac{x^2}{\sigma^2}\right)\left(\frac{x^2}{\sigma^2} - (\gamma^{1/2}-1)^2\right)}\mathbb{1}_{[\gamma^{1/2}-1,\gamma^{1/2}+1]}(x/\sigma)}_{\rho_\gamma\left(\frac{x}{\sigma}\right)}\,dx \\
&= \int \frac{1}{\sigma}\rho_\gamma\left(\frac{x}{\sigma}\right)dx = \Lambda
\end{aligned}$$

Lo que equivale a la expresión 1.3 probando la afirmación. □

Continuando ahora con el estudio de los valores propios y su distribución, se define la *densidad conjunta espectral* de $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ como sigue:

Definición 1.8 (**Densidad conjunta espectral**)

$$p(\vec{\lambda})d\vec{\lambda} = \frac{1}{Z_n}\left(\prod_{1\leq i,j\leq n}(\lambda_i - \lambda_j)\right)^{\beta}\prod_{i=1}^{n}\mathrm{e}^{-\frac{\lambda_i}{2}}d\lambda_i,$$

con Z_n una constante de normalización y $\beta = 1, 2$ para las familia GOE *y* GUE *respectivamente.*

Convencionalmente los probabilistas ordenan los valores propios de matrices aleatorias de dimensiones $n \times n$ decrecientemente, es decir, $\lambda_1^{(n)} \geq \lambda_2^{(n)} \cdots \geq \lambda_n^{(n)}$. Además, introducen conceptos como *bulk* y *edge* para referirse al conjunto de valores propios aglomerados y a los valores propios extremos $\lambda_1^{(n)}$ y $\lambda_n^{(n)}$, respectivamente.

Siguiendo esta línea, existen dos problemas de particular interés para los estudiosos de la teoría *RMT*. El primero tiene que ver con la convergencia de la medida espectral empírica, es decir, estudiar los casos en que existe una densidad límite. Mientras que el segundo, consiste en encontrar la distribución límite del *edge*, es decir, la distribución límite de los valores propios $\lambda_1^{(n)}$ y $\lambda_n^{(n)}$. Estos últimos, especialmente $\lambda_1^{(n)}$, han generado una literatura extensa al respecto, lográndose así, definir nuevas distribuciones que aparecen en diversos campos que a priori no están relacionados. Particularmente, se trata de leyes que aparecen en áreas como comunicación inalámbrica, rendimiento financiero, modelos de crecimiento estocástico, y hasta transiciones de fases en sistemas de partículas, como es el caso de [4] y [8] que constituyen ejemplos en los que se percibe la noción de *universalidad*.

1.3. Distribuciones Tracy-Widom

Antes de estudiar estas leyes universales se debe revisar un concepto previo que es necesario para su definición.

Definición 1.9 (**Función Airy**) *Se denomina función Airy, denotada como $Ai(x)$, a una de las soluciones de la ecuación diferencial*

$$y'' - zy = 0,$$

conocida justamente como ecuación de Airy y que aparece en el campo de la física, tanto en óptica como en mecánica cuántica. Su definición explícita es la que sigue:

$$Ai(x) = \frac{1}{\pi}\int_0^{\infty} cos\left(\frac{t^3}{3} + xt\right)dt$$

Craig Tracy y Harold Widom estudian en [19] la distribución del primer valor propio de matrices *GUE* en términos de un sistema de ecuaciones diferenciales parciales (edp) completamente integrable. La distribución resultante que obtuvieron se conoce como *Ley Tracy Widom* y ha sido por años objeto de interés para los campos mencionados anteriormente debido a su carácter universal. Finalmente, la definición formal de estas leyes surge luego de haber estudiado los sistemas de edp y las ecuaciones de Painlevé, obteniéndose:

Definición 1.10 (Distribuciones Tracy-Widom) *Se definen las siguientes:*

(i) ***Ley Tracy-Widom de orden 1:*** *También llamada ley Tracy-Widom GOE. Se define de acuerdo a su función de distribución que viene dada por:*

$$F_{\beta=1}(s) = exp\left\{-\frac{1}{2}\int_s^\infty q(x) + (x-s)q^2(x)dx\right\}, \quad s \in \mathbb{R},$$

donde la función $q(x)$ es la solución de la ecuación diferencial (no lineal) de Painlevé II

$$q''(x) = xq(x) + 2q^3(x)$$

$$q(x) \sim Ai(x) \text{ en el regimen } x \to +\infty$$

(ii) ***Ley Tracy-Widom de orden 2:*** *También llamada ley Tracy-Widom* GUE. *Se define de acuerdo a su función de distribución que viene dada por:*

$$F_{\beta=2}(s) = exp\left\{-\int_s^\infty (x-s)q^2(x)dx\right\}, \quad s \in \mathbb{R},$$

donde la función $q(x)$ sigue siendo la solución de la ecuación de Painlevé descrita arriba.

Observación: También existe la ley Tracy-Widom asociada al *GSE Ensemble*, cuya función de distribución $F_{\beta=4}$ no será definida en este trabajo por el mismo argumento por el cual no se definió la familia *GSE*.

La distribución Tracy-Widom tiene una forma bastante característica: es asimétrica, su cola izquierda decae como e^{-N^2} y la derecha como e^{-N} según se aprecia en la figura 1.1.[1] Además, la utilidad de definirla radica en que esta ley proporciona una forma explícita (bajo ciertas condiciones) para la distribución del primer valor propio de matrices aleatorias, lo cual permite probar resultados y propiedades, ya sea teórica o empíricamente.

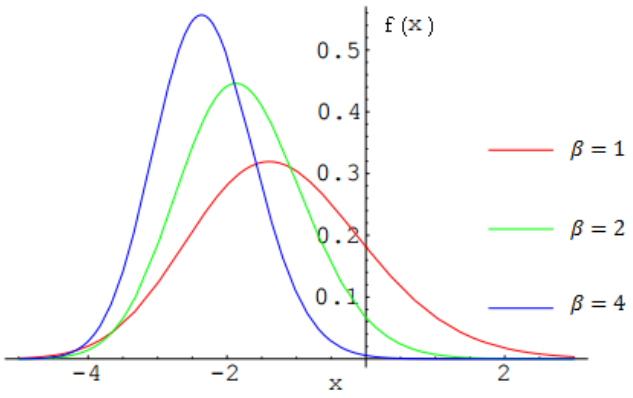


Figura 1.1: Densidad de probabilidad distribuciones Tracy-Widom asociadas a distintos β.

[1] Imagen extraída desde [20].

¿Cuáles son entonces estas condiciones y matrices especiales? Es la pregunta que responden los siguientes teoremas, cuyos enunciados han sido extraídos de [12].

Teorema 1.11 (**Ley Tracy-Widom *GOE***) *Sea X una matriz aleatoria de dimensiones $n \times p$, cuyas entradas reales $\{X_{ij}\}_{i\in[n],j\in[p]}$ se distribuyen $\mathcal{N}(0,1)$. Sea W su matriz de Wishart asociada, es decir, $W = XX'$ y sean $\lambda_1(W) \geq \cdots \geq \lambda_n(W)$ sus valores propios. Se definen :*

(i) $\mu_{np} = \left(\sqrt{p-1} + \sqrt{n}\right)^2$

(ii) $\sigma_{np} = \left(\sqrt{p-1} + \sqrt{n}\right)\left(\frac{1}{\sqrt{p-1}} + \frac{1}{\sqrt{n}}\right)^{1/3}$,

como constantes de centro y de escala respectivamente.

Entonces, si $n/p \to \gamma < 1$ se cumple:

$$\frac{\lambda_1(W) - \mu_{np}}{\sigma_{np}} \longrightarrow W_1 \sim F_1.$$

O equivalentemente se dice, $\sigma_{n,p}W_1 + \mu_{n,p}$ aproxima la distribución de $\lambda_1(W)$, donde W_1 tiene una función de distribución que corresponde a $F_{\beta=1}$.

Observación: El teorema sigue siendo válido cuando $n > p$ y $n, p \to \infty$ para la matriz de Wishart $X'X$ siempre y cuando se reviertan los roles de n y p en las definiciones de las constantes μ_{np} y σ_{np}.

Teorema 1.12 (**Ley Tracy-Widom GUE**) *Sea X una matriz aleatoria compleja de dimensiones $n \times p$, cuyas entradas $\{X_{ij}\}_{i\in[n],j\in[p]}$ se distribuyen $\mathcal{N}(0,1)$, es decir, para $1 \leq i \leq n, 1 \leq j \leq p$, se cumple que $\mathrm{Re}\,(X_{ij}), Im\,(X_{ij}) \sim \mathcal{N}(0, 1/2)$.*

Sea W su matriz de Wishart asociada, es decir $W = XX^*$, y sean $\lambda_1(W) \geq \cdots \geq \lambda_n(W)$ sus valores propios. Se definen :

- $\mu'_{np} = \left(\sqrt{p} + \sqrt{n}\right)^2$
- $\sigma'_{np} = \left(\sqrt{p} + \sqrt{n}\right)\left(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{n}}\right)^{1/3}$,

como constantes de centro y de escala respectivamente.

Entonces, si $n/p \to \gamma < 1$ se cumple:

$$\frac{\lambda_1(W) - \mu'_{np}}{\sigma'_{np}} \longrightarrow W_2 \sim F_2.$$

O equivalentemente se dice, $\sigma_{n,p}W_2 + \mu_{n,p}$ aproxima la distribución de $\lambda_1(W)$, donde W_2 tiene una función de distribución que corresponde a $F_{\beta=2}$.

Ejemplo 3 (**Aplicado al problema de detección de un target**) Siguiendo con el problema de detección de un *target*, estudiado en [10], y bajo el mismo escenario descrito en el Ejemplo 2 , se busca probar la siguiente afirmación:
"Para valores de N y M suficientemente grandes, con $N/M = \gamma \geq 1$, el primer valor singular $\sigma_1^{(M)}$ de la matriz A satisface

$$\sigma_1^{(M)} \sim \sigma \left[\gamma^{1/2} + 1 + \frac{1}{2M^{2/3}} \left(1 + \gamma^{-1/2}\right)^{1/3} W_2 + o\left(\frac{1}{M^{2/3}}\right) \right], \tag{1.4}$$

en distribución, donde W_2 sigue una distribución Tracy-Widom de orden 2"

Demostración. Lo primero que debe notarse es que como existen más receptores que emisores, es decir $N \geq M$, se tiene que $\frac{N}{M} = \gamma \geq 1$, por lo que el Teorema 1.12 es aplicable.

Recapitulando, se tiene que $A = \underbrace{A_0}_{\equiv 0} + W$, luego $A = W$ de dimensiones $N \times M$ y con entradas i.i.d. $\mathcal{N}(0, \sigma^2/M)$.

Con esto, si se denota como $\lambda_1(B) \geq ... \geq \lambda_M(B)$ a los valores propios de una matriz B cualquiera, se tendrá:

$$\lambda_1(AA^*) = \left(\sigma_1^{(M)}\right)^2.$$

Luego, aplicando el Teorema 1.12 :

$$\lambda_1(AA^*) = \left(\sigma_1^{(M)}\right)^2 \sim \frac{\sigma^2}{M} \left[\sigma_{M,N} W_2 + \mu_{M,N}\right],$$

donde el ponderador σ^2/M se debe a que la varianza de las entradas es σ^2/M y no 1 como en el enunciado estándar, y donde W_2 sigue una distribución Tracy-Widom de orden 2.

Y, reemplazando las constantes $\sigma_{M,N}$ y $\mu_{M,N}$:

$$\begin{aligned}
\left(\sigma_1^{(M)}\right)^2 &\sim \frac{\sigma^2}{M} \left[\left(\sqrt{M} + \sqrt{N}\right) \left(\frac{1}{\sqrt{M}} + \frac{1}{\sqrt{N}} \right)^{1/3} W_2 + \left(\sqrt{M} + \sqrt{N}\right)^2 \right] \\
&\sim \frac{\sigma^2}{M} \left[\sqrt{M} \left(1 + \frac{\sqrt{N}}{\sqrt{M}}\right) \left(\frac{1}{\sqrt{M}} \left(1 + \frac{\sqrt{M}}{\sqrt{N}}\right) \right)^{1/3} W_2 + \left(\sqrt{M} \left(1 + \frac{\sqrt{N}}{\sqrt{M}}\right) \right)^2 \right] \\
&\sim \frac{\sigma^2}{M} \left[\sqrt{M} \left(\gamma^{1/2} + 1\right) \frac{1}{\sqrt{M}^{1/3}} \left(1 + \gamma^{-1/2}\right)^{1/3} W_2 + M \left(1 + \gamma^{1/2}\right)^2 \right] \\
&\sim \frac{\sigma^2}{M} \left[M^{1/3} \left(\gamma^{1/2} + 1\right) \left(1 + \gamma^{-1/2}\right)^{1/3} W_2 + M \left(1 + \gamma^{1/2}\right)^2 \right] \\
&\sim \sigma^2 \left[\frac{1}{M^{2/3}} \left(\gamma^{1/2} + 1\right) \left(\gamma^{-1/2} + 1\right)^{1/3} W_2 + \left(\gamma^{1/2} + 1\right)^2 \right].
\end{aligned}$$

Denótese $c_1 = \sigma^2 \left(1 + \gamma^{1/2}\right)^2 \qquad c_2 = \dfrac{\sigma^2 \left(1 + \gamma^{1/2}\right) \left(1 + \gamma^{-1/2}\right)^{1/3}}{M^{2/3}}.$

Luego,

$$\left(\sigma_1^{(M)}\right)^2 \sim c_1 + c_2 W_2. \tag{1.5}$$

Considérese ahora el desarrollo en serie de Taylor de la función $f(x) = \sqrt{x}$ en torno a c_1. Esto es,

$$f(x) = f(c_1) + f'(c_1)(x - c_1) + \frac{f''(c_1)}{2}(x - c_1)^2 + \cdots + o(n)$$

es decir,

$$\sqrt{x} = \sqrt{c_1} + \frac{(x - c_1)}{2\sqrt{c_1}} + o\left(\frac{1}{c_1^{3/2}}\right). \tag{1.6}$$

Como $\sigma_1^{(M)} > 0$ se tiene, de acuerdo a la ecuación 1.6 , para $x = \left(\sigma_1^{(M)}\right)^2$ lo siguiente:

$$\sigma_1^{(M)} = \sqrt{\left(\sigma_1^{(M)}\right)^2} = \sqrt{c_1} + \frac{\left(\sigma_1^{(M)}\right)^2 - c_1}{2\sqrt{c_1}} + o\left(\frac{c_2}{2c_1^{3/2}}\right).$$

Finalmente, para concluir basta con conocer la distribución de $\left(\sigma_1^{(M)}\right)^2$, pues por 1.5 ,

$$\sigma_1^{(M)} \sim \sqrt{c_1} + \frac{c_1 + c_2 W_2 - c_1}{2\sqrt{c_1}} + o\left(\frac{c_2^2}{2c_1^{3/2}}\right).$$

Reemplazando las constantes c_1 y c_2 queda lo siguiente:

$$\begin{aligned}
\sigma_1^{(M)} &\sim \sigma(\gamma^{1/2} + 1) + \frac{\sigma^2(\gamma^{1/2} + 1)\left(\gamma^{-1/2} + 1\right)^{1/3} W_2}{2\sigma(\gamma^{1/2} + 1)M^{2/3}} + o\left(\frac{\sigma}{2M^{2/3}}\right) \\
&\sim \sigma\left[\gamma^{1/2} + 1 + \frac{(\gamma^{-1/2} + 1)^{1/3} W_2}{2M^{2/3}} + o\left(\frac{1}{2M^{2/3}}\right)\right].
\end{aligned}$$

Lo que equivale a la expresión 1.4 probando la afirmación. □

Capítulo 2

Estudio del Modelo de Factores

En el presente capítulo se discuten las circunstancias y motivaciones que dieron origen al trabajo aquí expuesto. En una primera instancia se introducirá el concepto de factor y su importancia desde un punto de vista estadístico. Además, se describirá el modelo de factores asociado, considerando sus características y mencionando dónde y para qué se usa. Luego, durante el capítulo, se irá dando paso a la teoría matemática que hay detrás del modelo, en particular cómo se relaciona con la *RMT*, y cómo las propiedades y resultados clásicos de la *RMT* pueden ser aplicados en este contexto. Finalmente, se desarrollará el test de hipótesis en sus versiones *GUE* y *GOE*, se describirá cómo el *GOE* constituye la base del procedimiento y se presentarán las conjeturas que le dan soporte a su funcionamiento.

La primera causa que motiva este trabajo de tesis consiste en estudiar el test en su versión *GOE* pues esta evita la división artificial de la muestra que realiza la versión original *GUE*. Con esto se busca dilucidar cuáles de las propiedades y características pertenecientes a la versión *GUE* pueden ser aplicadas también al caso *GOE*.

Sin embargo, más allá de la motivación matemática por estudiar el caso *GOE* y la *RMT* en general, el objetivo principal consiste en desarrollar una batería de herramientas matemáticas basada en fundamentos propios de matrices aleatorias y que permita obtener algún tipo de conocimiento aplicado.

Para lo anterior, es vital el estudio del trabajo previo realizado en [17] y [15], el cual sentará las bases de este trabajo.

El conocimiento aplicado que puede obtenerse a partir de este estudio es de índole estadística. Especificamente, una vez que se procesa el conocimiento de las matrices aleatorias, y en particular, del comportamiento de su espectro, es posible desarrollar un procedimiento que entrega determinada información sobre una muestra, cuyo conocimiento tiene gran valor estadístico.

De qué se trata dicha información y cuál es su importancia, son las interrogantes que se responderán en este capítulo, sin perjucio de describir los métodos y técnicas estadísticas más convencionales que se encargan igualmente de contestarlas. Todo esto, para luego describir

el modelo matemático utilizado y detallar la construcción del procedimiento que se presenta como una propuesta alternativa y no tan convencional.

2.1. Número de factores

Una de las características que comparte la estadística con la *RMT*, es el hecho de que ambas lidian con el análisis de datos de gran dimensión. En este contexto, existen métodos estadísticos de reducción de datos que buscan obtener la mayor cantidad de información a partir del menor número de variables. A estas variables se les denomina *componentes principales*, *componentes latentes* o *factores* según sea la técnica utilizada y los procedimientos que se llevan acabo para economizar variables en pos de conservar información.

Los métodos más convencionales que cumplen esta descripción son el análisis de componentes principales (o *PCA*) y el análisis factorial.

El análisis factorial es una técnica que se encarga de describir la variabilidad entre variables observables y posiblemente correlacionadas, en términos de unos pocos factores. Así, las variables observables pueden modelarse como una combinación lineal de los factores. Mientras que *PCA* es un procedimiento que permite transformar un número considerable de variables observables y correlacionadas, en un número menor de nuevas variables sin correlación.

Sin duda, independientemente de cuál sea el método utilizado, el conocimiento del número de estos factores es información que tiene gran valor estadístico, pues abre paso a realizar estudios de estimación o pronósticos. Además, el número de factores puede tener diversas interpretaciones económicas e importantes consecuencias teóricas, por lo su conocimiento es la valiosa información que se obtiene como producto del procedimiento.

Sin embargo, existen casos en los que no hay acuerdo sobre el número de factores entre economistas y estudiosos del tema. Esta falta de consenso podría deberse a la variedad de técnicas estadísticas, como las descritas, que son utilizadas para estimarlo. Ejemplos de este fenómeno son los datos macroeconómicos relacionados a series de tiempo en U.S.A y los datos financieros de retornos accionarios que fueron objeto de análisis en [18], [17] y [15]. En ellos no se logra una resolución explícita sobre el número de factores que rigen los datos.

En el afán de atacar la incerteza acerca del número de factores, surgen técnicas y métodos estadísticos alternativos basados en materias tangenciales a la estadística, tal como el que da origen a este trabajo y cuyo modelo se detalla en la siguiente sección.

2.2. Modelo de Factores

Los modelos de factores son ampliamente usados en macroeconomía y finanzas, y han despertado gran interés en el área de investigación, especialmente cuando se combinan con *LDDA*. En finanzas, constituyen la base del modelamiento de retornos accionarios bajo ciertas condiciones de arbitraje estudiadas en [6]. Mientras que en macroconomía son utilizados para monitorear la actividad económica, construcción de índices de inflación y análisis de política monetaria, entre otros. Su mayor utilidad radica en que son modelos que dan paso a realizar pronósticos, predicciones, evaluación, estimaciones o simplemente análisis general de temas o variables económicas de interés.

Las definiciones siguientes de estos modelos constituyen una versión simplificada de [9], pero no por ello menos precisas. Simplemente se trata de definiciones más accesibles y que encajan perfecto dentro del marco de este trabajo. Si aún así, a las definiciones se les quisiése dotar de toda la rigurosidad matemática, habría que referirse a [9], en donde las variables aleatorias se definen dentro de un espacio de Hilbert $L_2(\Omega, \mathcal{F}, \mathbb{P})$ siendo $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad dado.

Para introducir el modelo, primero es necesario definir uno más general estudiado en [9] , desde el cual se deriva.

Definición 2.1 (**Modelo Dinámico Generalizado de Factores**) *Matemáticamente, el modelo considera T observaciones $X_1, \ldots, X_T$ de vectores n-dimensionales correspondientes a series de tiempo de variables o indicadores financieros. Básicamente, el modelo se resume en:*

$$X_{it} = \Lambda_{i1}(L)F_{1t} + \cdots + \Lambda_{ik}(L)F_{kt} + \mathrm{e}_{it} \quad i = 1, \ldots, n, t = 1, \ldots, T.$$

O más compactamente,

$$X_t = \Lambda(L)F_t + \mathrm{e}_t \quad t = 1, \ldots, T, \tag{2.1}$$

donde

(i) $\Lambda(L)$ *es una matriz de dimensiones $n \times k$ cuyos términos $\{\Lambda_{ij}\}_{1\leq i\leq,1\leq j\leq k}$ corresponden a polinomios infinitos en el operador L (operador* lag*)*[1]*. Esto es:*

$$\Lambda_{ij}(L) = \sum_{u=0}^{\infty} \Lambda_{ij}^{(u)} L^u.$$

A los coeficientes $\Lambda_{ij}^{(u)}$ se les denomina ***factor loadings****.*

(ii) F_t *corresponde a un vector k-dimensional y representa a los* ***factores*** *en el tiempo t.*

(iii) e_t *corresponde a un vector n-dimensional estacionario y posiblemente correlacionado, cuyos términos se denominan* ***términos de idiosincrasia****.*

[1] En terminología usada en series de tiempo, el operador *lag* corresponde al operador que observa el tiempo desplazado

Observación: A la matriz e dada por los vectores columnas e_t también se le llama *matriz de ruido*, mientras que a $\Lambda(L)F_t$ se le denomina *parte sistemática.*

Sin embargo, existe un caso particular de este modelo en donde $\Lambda_{ij}(L)$ no depende de L para ningún $i = 1, \ldots, n$ ni $j = 1, \ldots, k$. Tal caso es introducido en [6] para el estudio de la teoría de arbitraje (o *APT* por *Arbitrage Pricing Theory*) y se le llama simplemente *modelo de factores.*

Definición 2.2 (**Modelo de Factores**) *Bajo la notación ya introducida, el modelo consiste en:*

$$X = \Lambda F + \mathrm{e}, \tag{2.2}$$

donde

(i) Λ *es una matriz determinista de dimensiones* $n \times k$ *a la que también se le denomina* ***factor loadings****.*

(ii) F *corresponde a los* ***factores*** *y tiene dimensiones* $k \times T$*.*

(iii) e *es la matriz de* ***términos de idiosincrasia****, tiene dimensiones* $n \times T$ *y sus filas admiten correlación.*

Para continuar con la descripción del modelo, sus supuestos y los resultados que a partir de él se obtienen, es necesario introducir cierta notación y conceptos previos.

Con respecto a la notación, a los k factores, vistos como vectores de observaciones, se les denota como F_t. Por su parte, para $n = 1, \ldots, N$, el vector $(\mathrm{e}_{1t}, \ldots, \mathrm{e}_{nt})'$ que corresponde a un vector de términos de idiosincrasia, se le denota como $\mathrm{e}_t(n)$.

Definición 2.3 *Con respecto a los conceptos, se describen rápidamente los siguientes:*

(i) Se define la función $c_{ij}(u)$ *de autocovarianzas de los términos de idiosincrasia como*

$$c_{ij}(u) := \mathbb{E}(\mathrm{e}_{i,t}\mathrm{e}_{j,t-u}) \quad \forall u \geq 0.$$

(ii) Sean $l_{1n} \geq \cdots \geq l_{nn}$ *los valores propios de* $\mathrm{e}_t(n)\mathrm{e}_t'(n)$ *para cada* $n = 1, \ldots, N$*. Se define su distribución espectral* $H_n(\lambda)$ *como:*

$$H_n(\lambda) = 1 - \frac{1}{n}|\{i \leq n : l_{in} > \lambda\}|.$$

Con esto, es posible enumerar los supuestos que, sumados a la definición de un estadístico particular, se enmarcan dentro de los principales resultados de matrices aleatorias que dan explicación al funcionamiento del test de hipótesis y al posterior procedimiento que se construye a partir de él.

S1: (i) Los factores F_t siguen un proceso de media cero y varianza no-degenerada[2].
(ii) Para cada n, el vector de observaciones $\mathrm{e}_t(n)$ es independiente de F_t.

S2: Los términos de idiosincrasia siguen un proceso estacionario gaussiano de media cero con $c_{ij}(u) = 0$ para todo $u \neq 0$.

S3: Sea $c_{n,T/2}$ una raíz en $[0, l_{1n}^{-1})$ de la ecuación:

$$\int \left(\frac{\lambda c_{n,T/2}}{1 - \lambda c_{n,T/2}} \right)^2 dH_n(\lambda) = \frac{T/2}{n}.$$

Entonces cuando $n, T \to \infty$ tal que $\gamma := n/T$ permanece en un compacto de $(0, \infty)$, se asume que $\limsup l_{1n} < \infty$, $\liminf l_{nn} > 0$ y $\limsup l_{1n} c_{T/2,n} < 1$.

S4: El k-ésimo valor propio de $\Lambda\Lambda'$ diverge a infinito más rápido que $n^{2/3}$.

Con el fin de generar intuición acerca de cómo operan los supuestos se señala lo siguiente:

- El supuesto S1 es un supuesto estándar y propio del modelo de factores.
- El supuesto S2 incide directamente en la base teórica que soporta al test, por lo que su función quedará explicitada más adelante.
- Por otra parte, los supuestos S3 y S4 tienen relación con el comportamiento asintótico de la matriz de covarianza de $\mathrm{e}_t(n)$.
- Con respecto al supuesto S3, la desigualdad $\limsup l_{1n} < \infty$ procura que el efecto de la matriz de ruido se mantenga acotado, mientras que la desigualdad $\liminf l_{nn} > 0$ procura que no se degenere. Por otra parte, la última desigualdad $\limsup l_{1n} c_{T/2,n} < 1$ procura que sus primeros valores propios no se dispersen demasiado, de manera que no sean malinterpretados como factores.
- Por último, el supuesto S4 apunta al fenómeno de separación del espectro (detallado más adelante) que es el principal responsable del funcionamiento del test.

2.3. Test *GUE*

Debido a la incertidumbre estadística que se presenta respecto al número de factores en la literatura en general, existe un constante esfuerzo por desarrollar técnicas que permitan cuantificar dicha incertidumbre. Allí es donde los tests de hipótesis son los protagonistas, pues determinan niveles de confianza para un número dado de factores.

La primera versión de los test de hipótesis de esta índole fue introducida en [15] y posteriormente formalizada en [17]. Esta última es la que se describe a continuación y la que da origen al estudio de su versión *GOE*.

[2]Existe además una condición sobre el cuarto momento que no se especificará aquí pues solo es relevante para efectos técnicos.

El objetivo consiste en testear la hipótesis de existencia de un cierto número dado de factores en una muestra que proviene del modelo 2.2. En otras palabras, lo que se testea en el fondo es el valor de la dimensión k dada en el modelo. Para ello, el test *GUE* se basa en argumentos provenientes de la teoría de matrices aleatorias, particularmente en el conocimiento de las distribuciones Tracy-Widom. Más en detalle, el test considera como hipótesis las siguientes:

$$H_0 : k = k_0 \text{ factores.}$$

$$H_1 : k_0 < k \leq k_1 \text{ factores.}$$

Donde k_1 es un número máximo de factores que se establece a priori. El test se desarrolla en los siguientes pasos:

(i) Dada la matriz de datos X, asumida como en el modelo 2.2 , se procede a dividirla temporalmente en dos periodos de igual largo, de modo que se construya una matriz compleja como sigue:

$$\hat{X}_j = X_j + iX_{j+\frac{T}{2}}. \tag{2.3}$$

(ii) Se determinan los valores propios $\hat{\gamma}_1 > \hat{\gamma}_2 > \ldots > \hat{\gamma}_l$ de la matriz de covarianza de X, es decir,

$$\tfrac{2}{T}\sum_{j=1}^{T/2} \hat{X}_j \hat{X}_j' \ .$$

(iii) Se calcula el estadístico

$$\hat{R} = \max_{k_0 < i \leq k_1} \frac{\hat{\gamma}_i - \hat{\gamma}_{i+1}}{\hat{\gamma}_{i+1} - \hat{\gamma}_{i+2}}. \tag{2.4}$$

(iv) Finalmente, se acepta o rechaza H_0 de acuerdo a la regla de decisión dada por la tabla 3.3 de valores críticos para el estadístico $\hat{R}$ que se halla en el anexo A . Estos valores críticos corresponden a los tamaños $15\,\%, 10\,\%, 9\,\%, \ldots, 1\,\%$ y son obtenidos mediante simulaciones numéricas de tipo Monte Carlo que se explicitarán más adelante. El tamaño escogido α será un parámetro que representa el nivel de significancia del test. Esta tabla tiene estrecha relación con la tabulación de valores de la distribución Tracy-Widom *GUE*, por argumentos que se desarrollarán en lo que sigue.

Con el test ya desarrollado, es posible dar paso al argumento teórico que lo sostiene, el cual viene de la mano de la *RMT*. En primer lugar, se debe extender el trabajo de [19] a más dimensiones y definir una ley Tracy-Widom conjunta, lo cual hará posible presentar los principales resultados, descritos en [17], en los que se basa el test *GUE*.

Cuando Tracy y Widom estudiaron las matrices *GUE*, no quedaron conformes al definir las leyes Tracy-Widom, sino que además procuraron extender este resultado al estudio de los m primeros valores propios de una matriz *GUE*. Intuitivamente, se extiende este resultado como sigue:

Definición 2.4 (Distribución Tracy-Widom conjunta) *Sea X una matriz GUE de $N \times N$ y denótese como $d_1 \geq ... \geq d_N$ a sus valores propios. Luego $d_i' := N^{2/3}(d_i - 2)$, con $i = 1, \ldots, N$, corresponde a los mismos valores propios adecuadamente escalados y centrados. Se define la distribución Tracy-Widom conjunta N-dimensional de orden 2 como la distribución conjunta de $d_1', ..., d_N'$.*

Ahora bien, una vez definida la distribución Tracy-Widom conjunta, es posible explicar la base del test que consiste en la idea de separación del espectro. Esto es, cuando k es el verdadero número de factores, los primeros k valores propios se separan del *bulk* saliéndose de escala. Este argumento se fundamenta en las siguientes tres observaciones:

(i) La primera es que, dado que se trabaja con el caso en que la matriz Λ no depende de L, se tiene que la matriz de datos X posee la estructura de factores en el sentido en que se describe en [6]. En otras palabras, X se ajusta al modelo 2.2 .

(ii) La segunda consiste en que al considerar el supuesto S2 se tiene como consecuencia que la matriz de covarianza

$$S := \frac{1}{T} \sum_{t=1}^{T} \mathrm{e}_t(n) \mathrm{e}_t'(n)$$

converge en distribución a una matriz Wishart. Luego, como la matriz X posee una estructura de k factores, los valores propios $\gamma_1, \ldots, \gamma_k$ divergen cuando $n, T \to \infty$, mientras que el resto de ellos $\gamma_{k+1}, \ldots, \gamma_n$ se aproxima a los valores propios de la matriz de covarianza S, cuya distribución converge a la de la matriz Wishart en cuestión.

(iii) Finalmente, la tercera observación indica que en el régimen asintótico $n, T \to \infty$, bajo H_1, se tiene que $k_0 < k$, lo que implica que

$$R \geq \frac{\gamma_k - \gamma_{k+1}}{\gamma_{k+1} - \gamma_{k+2}}$$

Luego, dado que $\gamma_k \to \infty$ y que γ_{k+1} y γ_{k+2} permanecen acotados por aproximarse a los valores propios de S, se tiene que el estadístico $\hat{R}$ diverge bajo la hipótesis alternativa. Sin embargo, bajo la hipótesis nula $\hat{R}$ se aproxima bien en distribución a

$$\max_{0 < i \leq k_1 - k_0} \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_{i+2}},$$

donde λ_i es el i-ésimo valor propio de la matriz Wishart.

La distribución marginal del primer valor propio de matrices Wishart fue estudiada en [13] en su versión más simple y luego extendida a más valores propios y a casos de matrices Wishart singulares en [16]. La versión más definitiva de este resultado (enunciado y probado en [16]) se resume como sigue:

Lema 2.5 *En el regimen asintótico $n, T \to \infty$ en el que $\gamma = n/T$ permanece en un compacto de $(0, \infty)$, se tiene que para cualquier entero positivo r, la distribución conjunta de los valores propios centrados y escalados, $\sigma_{n,T}^{-1}(\lambda_1 - \mu_{n,T}), \ldots, \sigma_{n,T}^{-1}(\lambda_r - \mu_{n,T})$, de la matriz de Wishart S, converge débil a la distribución Tracy-Widom conjunta r-dimensional de orden 2.*

Observación: Las constantes de escala $\sigma_{n,T}$ y $\mu_{n,T}$ se definen de acuerdo a $c_{n,T/2}$ y a la distribución espectral H_n. Sus definiciones explícitas pueden encontrarse en [17] y no son presentadas aquí, pues el test *GUE* no las requiere debido a que en el estadístico se simplifican convenientemente.

En resumen, con el lema 2.5 se puede aventurar que, debido a que la distribución conjunta de $\gamma_{k+1}, \ldots, \gamma_{k+r}$ es aproximada por la de $\lambda_1, \ldots, \lambda_r$, entonces es posible testear la hipótesis nula de $k = k_0$ chequeando si los valores propios $\gamma_{k_0+1}, \ldots, \gamma_{k_0+r}$ provienen de una distribución Tracy-Widom conjunta de orden 2. Esto es lo que sugiere el siguiente teorema:

Teorema 2.6 *Bajo los supuestos S1, S2, S3 y S4, y bajo el regimen asintótico en que $n, T \to \infty$ de manera que $\gamma = n/T$ pertenece a un compacto de $(0, \infty)$, se tiene que, para cualquier entero positivo r, la distribución conjunta de los valores propios centrados y escalados, $\sigma_{n,T}^{-1}(\gamma_{k+1} - \mu_{n,T}), \ldots, \sigma_{n,T}^{-1}(\gamma_{k+r} - \mu_{n,T})$, converge débil a la distribución Tracy-Widom r-dimensional de orden 2.*

Observación: El Teorema 2.6 requiere mencionar comentarios:

(i) La misma observación que se hizo en el Lema 2.5 atañe a este resultado, es decir, las constantes solo se mencionan para dar formalidad al enunciado del teorema, dado que en la práctica no es necesario conocerlas.

(ii) La demostración de este teorema usa el Lema 2.5 y puede hallarse en [17] .

Finalmente, el siguiente teorema resume todas las propiedades del test *GUE*:

Teorema 2.7 *Bajo las condiciones del teorema 2.6 se tiene lo siguiente:*

(i) Bajo H_0, es decir, cuando $k = k_0$, el estadístico $\hat{R}$ converge en distribución al

$$\max_{0<i\leq k_1-k_0} \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_{i+2}}$$

donde $\lambda_1, \ldots, \lambda_{k_1-k_0+2}$ son v.a. con distribución Tracy-Widom (k_1-k_0-2)-dimensional de orden 2.

(ii) En contraste, bajo H_1, es decir, cuando $k_0 < k \leq k_1$, el estadístico $\hat{R}$ diverge a infinito en probabilidad.

Por lo tanto, el test GUE es consistente y tiene el tamaño correcto asintóticamente.

Observación: Dado el Teorema 2.7 es necesario señalar dos comentarios:

(i) El motivo por el cual la tabla 3.3 de valores críticos *GUE* se relaciona estrechamente con la distribución Tracy-Widom de orden 2, se debe al punto *(i)* de este teorema.

(ii) Dado que el teorema formaliza y resume los resultados ya descritos, para probarlo se utilizan argumentos propios de la demostración del Teorema 2.6 .

2.3.1. Determinando el número de Factores

Si bien el test es presentado como medio para contrastar hipótesis, también puede ser utilizado para estimar el número de factores. En este apartado se detalla el procedimiento que hace uso del test *GUE* para encontrar el número estimado de factores de una muestra.

El procedimiento básicamente consiste en aplicar de manera recursiva el test *GUE* contrastando en cada iteración un número distinto de factores. El procedimiento se detendrá cuando ya no sea capaz de rechazar hipótesis, en cuyo caso, dicha hipótesis brindará el número de factores buscado.

Más en detalle, supóngase que se sabe que el verdadero número de factores k se encuentra entre $\underline{k}$ y $\overline{k}$, conocidos a priori. Luego, se debe testear $H_0 : k = \underline{k}$ versus $H_1 : \underline{k} < k \leq \overline{k}$ con un nivel de significancia α. Si H_0 no es rechazada, se detiene, el número estimado de factores es $\underline{k}$. De lo contrario, si H_0 es rechazada, se debe testear $H_0 : k = \underline{k} + 1$ versus $H_1 : \underline{k} + 1 < k \leq \overline{k}$. Finalmente, se debe repetir el procedimiento hasta que H_0 no sea rechazada y considerar el número correspondiente de factores como el estimado.

Con este procedimiento, el número estimado de factores se acercará al verdadero número con probabilidad $1 - \alpha$ en el regimen $n \to \infty$.

2.4. Test *GOE*

En esta sección comienza formalmente a describirse la innovación realizada en el contexto del trabajo de tesis. La primera instancia a definir es cómo se diferencia la versión *GOE* de la original *GUE* ya descrita. Esta versión *GOE*, como es de esperar, también definirá un procedimiento que permitirá estimar el número de factores.

El test en su modalidad *GOE* surge debido al cuestionamiento al cual es sometida la división de la matriz X en la expresión 2.3 . Y la principal conclusión que se obtiene a partir de tal cuestionamiento, es que 2.3 constituye una forma práctica de construir una matriz compleja y que, por lo tanto, cualquier otra forma es bien recibida. Por este motivo, es razonable aventurar que no efectuar la división es otra forma factible -y más simple- de llevar a cabo el test, siempre y cuando se adapte su desarrollo al manejo de una matriz real y no compleja. Aquí es donde se hace alusión a la familia de matrices *GOE*.

El test *GOE*, como propuesta para evitar la división de la matriz X, tiene como hipótesis las mismas que el test *GUE* y se desarrolla como sigue:

(i) Dada la matriz X, con entradas reales y asumida como en el modelo 2.2, se determinan los valores propios $\gamma_1 > \gamma_2 > \ldots > \gamma_l$ de su matriz de covarianza ,

$$\frac{1}{T} \sum_{j=1}^{T} X_j X_j' \ .$$

(ii) Se calcula el estadístico

$$R = \max_{k0<i\leq k_1} \frac{\gamma_i - \gamma_{i+1}}{\gamma_{i+1} - \gamma_{i+2}} \tag{2.5}$$

(iii) Finalmente, se acepta o rechaza H_0 de acuerdo a la regla de decisión dada por la tabla 3.4 de valores críticos para el estadístico R. Ellos corresponden a los tamaños $15\,\%, 10\,\%, 9\,\%, \ldots, 1\,\%$ y se obtienen mediante simulaciones numéricas de tipo Monte Carlo que se detallarán en el capítulo siguiente. El tamaño escogido α será el nivel de significancia del test. Esta tabla, como es de esperar, tiene estrecha relación con la tabulación de valores de la distribución Tracy-Widom *GOE*, por argumentos análogos a los que se dieron en el caso *GUE*.

Análogamente, se presentan los resultados que constituyen la base para el funcionamiento del test *GOE* de la mano de la *RMT*, pero ahora en modalidad de dos conjeturas tipo resumen:

Conjetura 1 Asuma que los términos de idiosincrasia son variables aleatorias reales y que se cumplen los supuestos S1, S2, S3 y S4. En el regimen asintótico $n, T \to \infty$ tal que $\gamma = n/T$ permanece en un compacto de $(0, \infty)$, para cada entero positivo r, se tiene que la distribución conjunta de los valores propios centrados y escalados,

$$\sigma_{n,T}^{-1}(\lambda_1 - \mu_{n,T}), \ldots, \sigma_{n,T}^{-1}(\lambda_r - \mu_{n,T}),$$

de la matriz S converge a la distribución conjunta de

$$d_1' := N^{2/3}(d_1 - 2), \ldots, d_r' := N^{2/3}(d_r - 2),$$

donde los d_i son los valores propios de una matriz *GOE* de NxN. Como es de esperar esta distribución conjunta es la que se conoce como *distribución Tracy-Widom conjunta r-dimensional de orden 1.*

Esta conjetura ha sido apoyada por simulaciones numéricas en [13], y en esta línea también se tiene:

Conjetura 2 En el contexto de la conjetura 1 , se tiene:

(i) Bajo H_0, es decir, cuando $k = k_0$, el estadístico R converge en distribución al

$$\max_{0<i\leq k_1 - k_0} \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_{i+2}},$$

donde $\lambda_1, \ldots, \lambda_{k_1 - k_0 - 2}$ son v.a. con distribución Tracy-Widom $(k_1 - k_0 - 2)$-dimensional de orden 1.

(ii) En contraste, bajo H_1, es decir, cuando $k_0 < k \leq k_1$, el estadístico R diverge a infinito en probabilidad.

Capítulo 3

Análisis estadístico aplicado al Modelo de Factores

A continuación en el capítulo se describirán métodos, estudios y pruebas empíricas llevadas a cabo sobre el test *GOE* como forma de validarlo, conocer sus propiedades y evaluar su utilidad. Dicha evaluación se realiza en términos de los errores de tipo I y II y de algunas medidas asociadas. Todo esto para llegar a definir el *Procedimiento de factores* que se basa en la aplicación del test para entregar un número estimado de factores. Luego, se estudia el comportamiento y las características del procedimiento para luego evaluar su rendimiento acorde a las estimaciones que realiza y a los errores que comete. Por último, se describe una posible aplicación del *Procedimiento de factores* sobre datos reales.

3.1. Estudio usando el método de Monte Carlo

De acuerdo a la conjetura 2, en régimen asintótico, el estadístico del test converge, bajo la hipótesis nula, en distribución a

$$\max_{0<i\leq k_1-k_0} \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_{i+2}}$$

donde $\lambda_1, \ldots, \lambda_{k_1-k_0+2}$ son v.a. con distribución Tracy-Widom $(k_1 - k_0 + 2)$-dimensional de orden 1.

Esto implica que, conociendo las distribuciones de los λ_i es posible conocer la distribución de

$$\max_{0<i\leq k_1-k_0} \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_{i+2}},$$

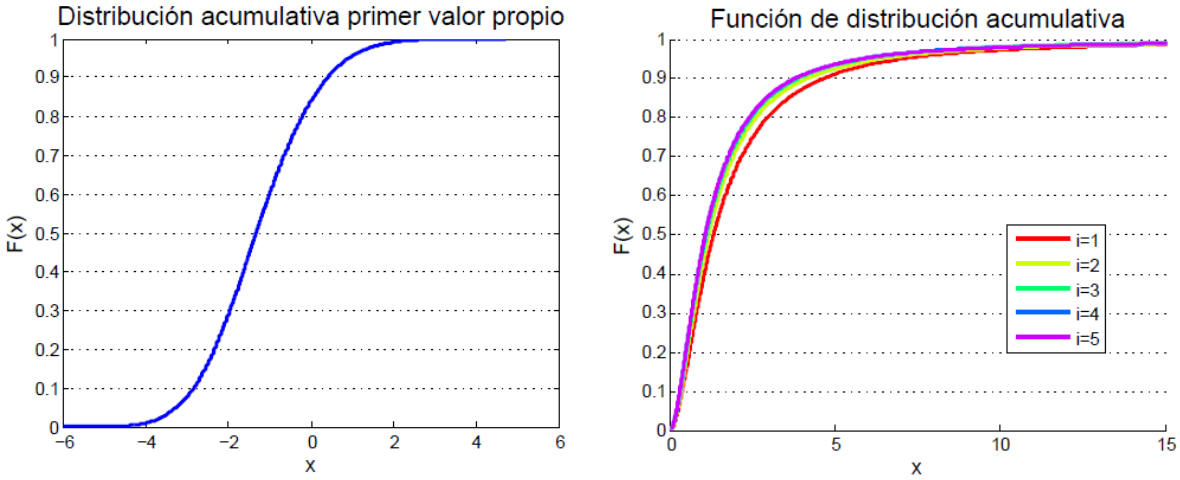
y por ende la de R.

Luego, el principal objetivo de esta sección es presentar un método que dé a conocer empíricamente las distribuciones de los λ_i. Esto es posible gracias a simulaciones del método de Monte Carlo, con las cuales se logra aproximar empíricamente la distribución Tracy-Widom

r-dimensional para $r \in \mathbb{N}$. Finalmente, con ella es posible calcular la distribución de R para luego tabular los valores críticos que formarán la regla de decisión del test.

Para llevar a cabo el objetivo, el método de Monte Carlo aproxima la ley Tracy-Widom *GOE* mediante el cálculo numérico de los valores propios de una matriz *GOE*. Es decir, aproxima la distribución Tracy-Widom 20-dimensional por la distribución conjunta de los primeros 20 valores propios de matrices pertenecientes a la familia *GOE*.

Más en detalle, se generan $N = 30000$ matrices *GOE* independientes y de dimensiones 1000×1000, de las cuales se obtienen los primeros 20 valores propios adecuadamente escalados y centrados, lo que genera 20 distribuciones de valores propios. Para el caso del primer valor propio, es decir, para la distribución Tracy-Widom *GOE* se observa, a la izquierda de la figura 3.1 , su función de distribución acumulativa. Mientras tanto, a la derecha se muestra, solo para ilustrar, la función de distribución acumulativa de $\frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_{i+2}}$ para $i = 1, \ldots, 5$.



(a) Distribución Acumulativa de λ_1

(b) Distribución Acumulativa de $\frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_{i+2}}$

Figura 3.1: Funciones de distribución acumulativa

Luego, a partir de las distribuciones de los valores propios λ_i, se deriva la distribución del estadístico R, de la cual se obtienen ciertos percentiles que más tarde se traducirán en valores críticos de distintos niveles de significancia del test.

Para efectos de simulaciones Monte Carlo se obtienen los percentiles $85\,\%, 90\,\%, 91\,\%, \ldots, 99\,\%$, con los que se hace posible tabular valores críticos para los niveles de significancia complementarios asociados $15\,\%, 10\,\%, 9\,\%, \ldots, 1\,\%$ respectivamente. La tabla generada se puede observar en el anexo B (ver tabla 3.4).

Por otra parte, el mismo ejercicio puede hacerse para el caso *GUE*. Es decir, es posible aproximar la ley Tracy-Widom *GUE* multidimensional por medio de las distribuciones de los valores propios de matrices *GUE*. Con ellas, también es factible tabular, tal como se describió mediante percentiles, los valores críticos, pero esta vez para el estadístico $\hat{R}$. Una versión de esta tabla para los niveles de significancia $15\,\%, 10\,\%, 9\,\%, \ldots, 1\,\%$ se encuentra en [17].

A modo de referencia se corren simulaciones Monte Carlo para $N = 30000$ matrices GUE con el fin de tabular los valores críticos para el estadístico complejo $\hat{R}$. Los objetivos de generar esta tabulación son:

En primer lugar, contrastar con la primera versión de la tabla de [17], cuya aparición se ha constituido en una referencia para este trabajo. En segundo lugar, contar con la tabla de valores críticos GUE permite desarrollar e implementar el test GUE a la par con el GOE, de modo que se hace factible realizar evaluaciones comparativas entre ambos.

Este último punto es parte también de las interrogantes matemáticas que motivan el desarrollo de este trabajo. Es decir, desde el punto de vista matemático, en particular, desde la RMT, es interesante conocer cómo se comparan ambos tests. Es claro que el punto desde el cual difieren se origina en la división de la muestra dada por la expresión 2.3. Luego, debido a esa diferencia, es posible que el resultado de las evaluaciones respectivas se incline favorablemente hacia el test GOE, ya que, al no realizar dicha división, este podría ser de mayor utilidad que el GUE. Sin embargo, no existe ninguna pista o argumento teórico que dé apoyo a este razonamiento, por lo que el resultado podría ser justamente el escenario opuesto, o incluso aquel en que la evaluación de ambos no difiera en lo absoluto.

La sección siguiente se encarga justamente de dar respuestas a esta pregunta mediante el desarrollo e implementación de ambos tests.

3.2. Implementación y evaluación del test GOE

Con la construcción de la tabla de valores críticos para ambos tests se da paso a su implementación utilizando un valor determinado del parámetro α. Este representa el nivel de signficancia del test en cuestión y se le denomina *tamaño nominal* cuando se refiere expresamente al parámetro de la implementación.

La implementación se realiza para ambos test, pero solo se ilustra para la versión GOE en el anexo C (ver código de la función $TestGOE$), pues la de GUE es análoga, salvo la división de la muestra y la tabla respectiva de valores críticos.

Después de la implementación, la primera evaluación que debe hacerse de un test de hipótesis es en términos de los errores de tipo I y II y de la potencia.

Definición 3.1 (**Errores test de hipótesis**) *Se definen los errores de tipo I y II de un test de hipótesis como sigue:*

- ***Error de Tipo I:*** *Se denota por α y se le llama también* ***falso positivo****.*

$$\alpha := \mathbb{P}(\textit{escoger } H_1 | H_0 \textit{ es cierta })$$

- ***Error de Tipo II:*** *Se denota por β y se le llama también* ***falso negativo****.*

$$\beta := \mathbb{P}(\textit{escoger } H_0 | H_1 \textit{ es cierta })$$

Definición 3.2 (**Potencia del test de hipótesis**) *Se define potencia del test como* $1-\beta$*, o alternativamente como:*

$$Potencia \ := \mathbb{P}(escoger \ H_1 | H_1 \ es \ cierta \ )$$

A continuación se detallan las pruebas y estudios empíricos que son parte de la evaluación. En primer lugar, se describen los experimentos relacionados al error de tipo I, para luego proseguir con la potencia.

3.2.1. Error de tipo I

Para estudiar el error de tipo I, el escenario a considerar como verdadero, es la existencia de k factores, lo que trae como consecuencia inmediata el fenómeno de separación del espectro de acuerdo al teorema 2.7 para el caso *GUE* y a la conjetura 2 para el *GOE*.

Observación: El entendimiento del fenómeno de separación del espectro podría llevar a concluir que la mera observación de los valores propios de la matriz de covarianza es suficiente para determinar el número k. Es decir, bastaría con examinar la cantidad de valores propios que se despegan del *bulk* para estimar el número de factores. Si bien esto es cierto, los métodos de inspección visual carecen de la rigurosidad estadística que permite cuantificar la incertidumbre asociada. Por esta razón, los tests de hipótesis y los procedimientos relacionados pueden ser interpretados como una formalización de los métodos empíricos más visuales.

Luego, basándose en dicho fenómeno de separación, se tiene que en el regimen asintótico bajo la hipótesis nula, $\gamma_1, \ldots, \gamma_k \to \infty$, mientras que el resto de los valores propios se mantienen acotados dentro del *bulk*. Esto implica que cuando la hipótesis nula es verdadera, es decir, cuando se testea $H_0 : k_0 = k$, el estadístico R se mantiene acotado. Como consecuencia, es de esperar que su valor no supere el valor crítico correspondiente y caiga dentro de la zona de aceptación del test el $(100-\alpha)\,\%$ de las veces, y por ende, rechace cuando no debe el $\alpha\,\%$ de las veces.

En otras palabras, es esperable que el comportamiento empírico del error de tipo I tenga estrecha relación con el tamaño nominal. Esto es lo que efectivamente ocurre tanto para el test *GUE* como para el *GOE*. El cálculo empírico que se realiza para justificar esto último consiste en contabilizar, de un total de N veces que se aplica el test, las oportunidades en que se equivoca y rechaza H_0 cuando el número real de factores es k. Para valores suficientemente grandes de N, el cálculo arroja una tasa α_{emp} que se asocia con la medida empírica del error de tipo I.

Para ilustrar lo anterior se generan muestras con 3 factores y se testea la hipótesis nula de $H_0 : k = 3$ factores versus una alternativa de $H_1 : 3 < k \leq 10$. Este ejercicio se realiza para tamaños nominales de $1\,\%, 5\,\%$ y $10\,\%$ y se hace en función de la cantidad de muestras N a las que se aplica el test. El resultado se muestra en la gráfica 3.2 .

Lo que se muestra en la figura 3.2 corresponde a las tasas empíricas obtenidas de las

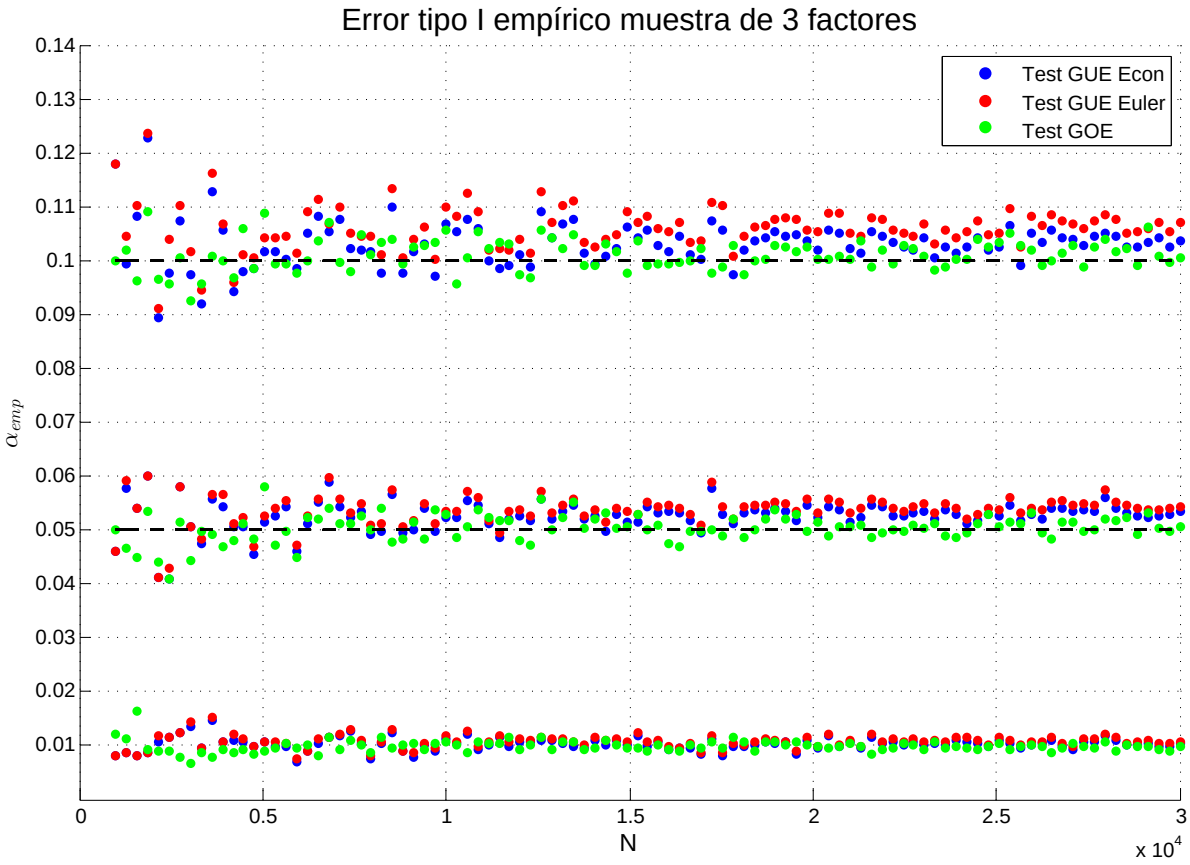


Figura 3.2: Error empírico de tipo I en función del número de muestras.

realizaciones del experimento. Las tendencias son los valores 0.01, 0.05 y 0.10, lo que evidencia la correspondencia con los tamaños nominales respectivos con los que se implementan los tests. Además, en la figura se muestra un código de colores[1] que representa cada versión del test. Esto es:

- En verde se grafica la implementación del test *GOE*;
- En rojo, la implementación del test *GUE* que utiliza la tabla de valores críticos 3.3 generada en el contexto de este trabajo. A este test se le llama *Test GUE Euler* debido al nombre que se la da al servidor[2] en el que se realizaron las simulaciones de Monte Carlo;
- En azul, la implementación del test *GUE* que utiliza la primera versión de la tabla de valores críticos publicada en [17]. A este test se le llama *Test GUE Econ* debido al nombre de la revista[3] que publicó el artículo.

En resumen, la gráfica mostrada entrega evidencia de que, al menos en términos de error de tipo I empírico, no existe gran diferencia entre un test y el otro.

En un segundo experimento relacionado al error de tipo I empírico para ambos tests, se realizan dos pruebas que involucran al resto de los tamaños nominales.

La primera de ellas cuantifica la diferencia entre el error de tipo I empírico y el tamaño nominal respectivo. Para ello se define una medida de discrepancia absoluta en función del tamaño. Por otra parte, para la segunda prueba se propone una medida que da cuenta de la diferencia relativa entre estos valores.

Definición 3.3 (**Discrepancia absoluta**) *Se define la medida de discrepancia absoluta entre error de tipo I empírico α_{emp} y el tamaño nominal α como sigue:*

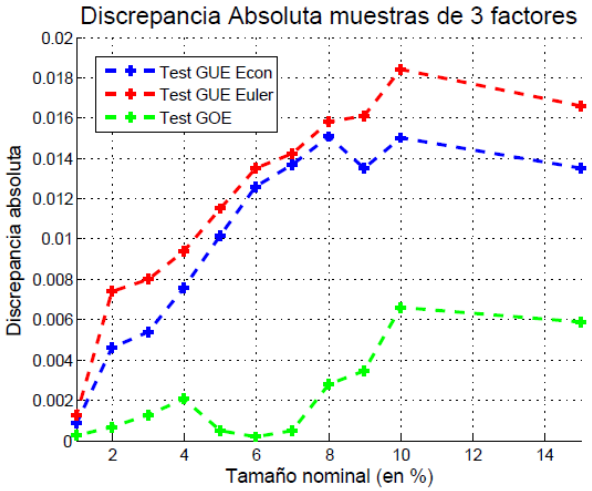
$$\textit{Discrepancia Absoluta} = |\alpha_{emp} - \alpha|$$

Luego, poniendo en práctica el mismo ejercicio que calcula la tasa α_{emp} para muestras generadas con 3 y 4 factores, se ilustra a continuación, la discrepancia absoluta de estas tasas con respecto a los tamaños nominales $1\,\%, 2\,\%, \ldots, 10\,\%$ y $15\,\%$ para los tests de nulas $H_0 : k = 3$ y $H_0 : k = 4$ factores respectivamente.

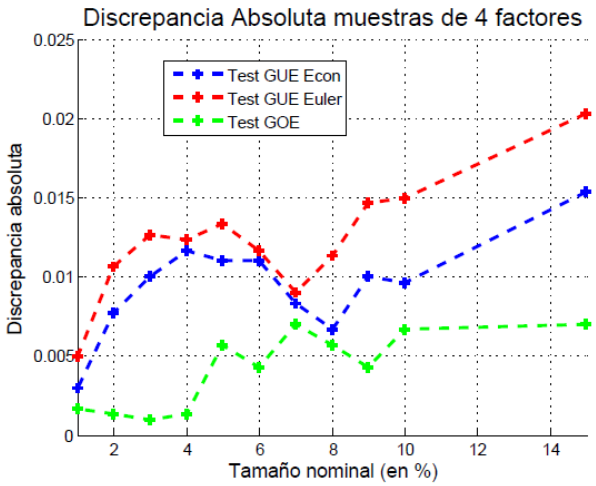
[1] Este código de colores se mantiene en las gráficas a continuación para efectos de comparación entre los tests.

[2] El servidor es propiedad del Departamento de Ingeniería Matemática de la Universidad de Chile.

[3] Abreviación del nombre de la revista Econometrica



(a) Test de $H_0 : k = 3$ factores



(b) Test de $H_0 : k = 4$ factores

Figura 3.3: Discrepancia absoluta de tamaños

La gráfica de la figura 3.3 se obtiene a partir de $N = 10000$ muestras generadas con 3 y 4 factores, y a modo de conclusión, se observa que en ambos casos de la figura, el test *GOE* posee una discrepancia menor que la de ambas versiones del test *GUE*.

Finalmente, la segunda prueba consiste en el cálculo de una medida que cuantifica la diferencia relativa entre la tasa α_{emp} y los tamaños nominales $1\,\%, \ldots, 10\,\%$ y $15\,\%$ en promedio. Para ello, se propone la medida de *discrepancia media relativa* que se define explícitamente como sigue:

Definición 3.4 (Discrepancia media relativa)

$$\textit{Discrepancia media relativa} \; = \frac{1}{11} \sum_{j=1}^{11} \frac{|\alpha_{emp}^{j} - \alpha^{j}|}{\alpha^{j}}$$

donde α_{emp} es la tasa empírica del error de tipo I como ya se había mencionado antes, α corresponde al tamaño nominal, y el superíndice j denota los 11 tamaños involucrados, a saber: $1\,\%, \ldots, 10\,\%$ y $15\,\%$.

Entonces, a partir de la generación de N muestras de 3 factores y de repetir una última vez el procedimiento descrito para el cálculo de la tasa α_{emp}, es posible graficar esta medida. Esta última se muestra como función de la cantidad N de muestras sobre las que se aplican los respectivos tests *GUE* y *GOE* de hipótesis nulas $H_0 : k = 3$ factores versus $H_1 : k_0 < k \leq k_1$.

Para hacer más cómoda la visualización, en la figura 3.4 solo se muestran los tests *GOE* y *GUE Euler*. Además, se muestra mediante una línea continua la tendencia central que siguen los puntos de cada test. Se destaca en la figura que el test *GOE* es propenso a tener menor discrepancia cuando se le compara con su versión *GUE*.

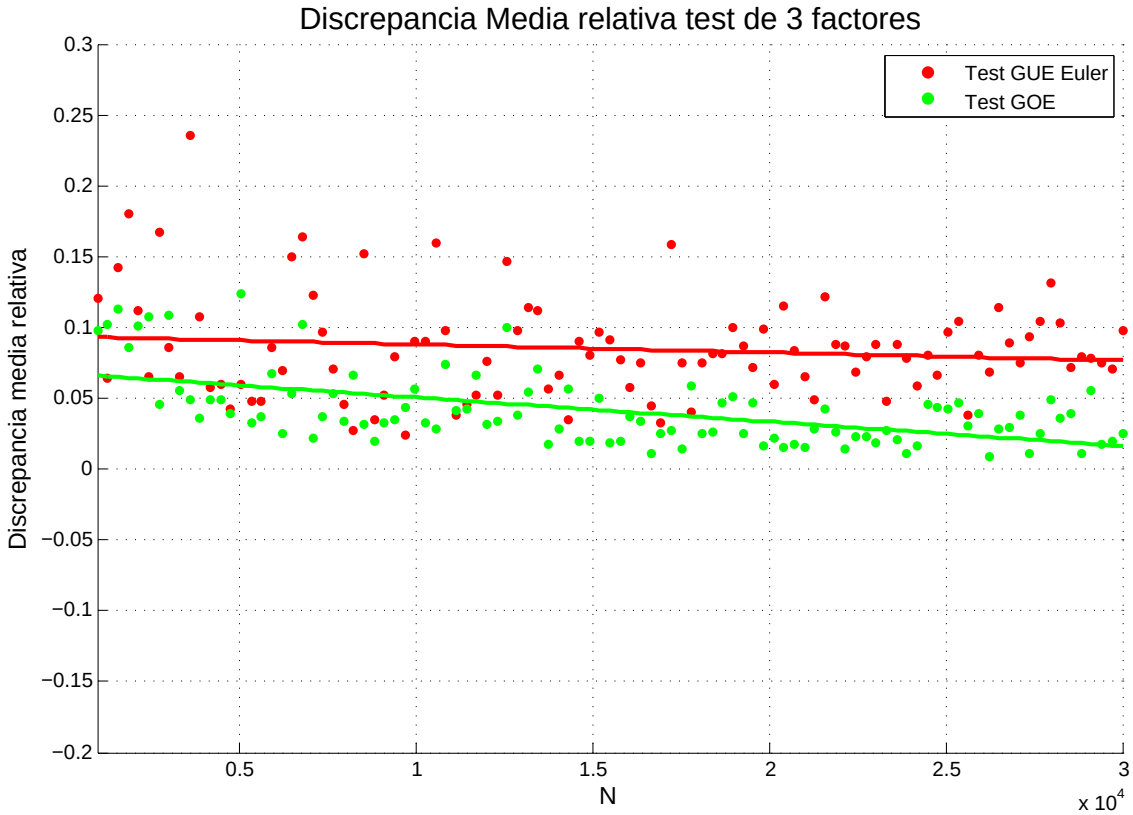


Figura 3.4: Discrepancia media relativa en función de N.

De acuerdo a los resultados obtenidos de las tres pruebas anteriores, se concluye que el test *GOE* no presenta grandes diferencias con respecto a su versión original *GUE*. No obstante, las pruebas también muestran evidencia a favor del *GOE*, por lo cual este se enmarca como un test de hipótesis idóneo para delimitar la incerteza estadística que se genera con respecto al número de factores.

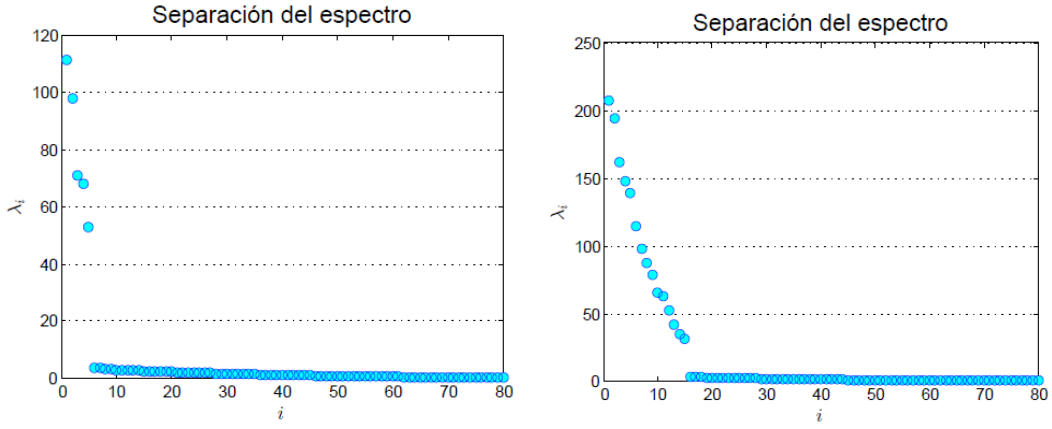
3.2.2. Potencia del test *GOE*

Una vez analizado todo lo relacionado a error de tipo I del test *GOE*, en una segunda instancia, se procede a estudiar el error de tipo II o complementariamente la potencia.

Para un test de hipótesis cualquiera, lo ideal sería tener errores de tipo I y II pequeños, aunque, es razonable considerar que obtener un error de tipo I pequeño conlleve a incrementos del error de tipo II o, equivalentemente, a una disminución en la potencia.

Sin embargo, este no es el caso. El test *GOE* presenta una particularidad en términos de errores de tipo I y II, a saber: presenta una especie de asimetría entre ambos que logra evidenciarse empíricamente. Mientras el error empírico de tipo I ronda el 5 % de acuerdo al tamaño nominal, el error de tipo II es muy cercano a cero a partir de un cierto valor $\alpha_0 > 0$[4].

La razón por la cual β adopta este comportamiento está implícita en el funcionamiento del test y de su estadístico R bajo la hipótesis alternativa. Para obtener el error de tipo II se debe tener en mente el escenario en el cual el número verdadero de factores k es tal que $k_0 < k \leq k_1$. En este caso, ya es sabido que $\gamma_1, \ldots, \gamma_k \to \infty$ cuando $n \to \infty$, escapándose del *bulk* y particularmente saliéndose de escala. Este fenómeno tiene incluso consecuencias visuales, pues, al ilustrar los valores propios, los primeros k son macroscópicamente visibles, mientras el *bulk* se observa como una nube de puntos indistinguibles. Este es el ya conocido efecto de separación del espectro y puede observarse a continuación en la figura 3.5 para muestras de 5 y 15 factores.



(a) Muestra de 5 factores (b) Muestra de 15 factores

Figura 3.5: Fenómeno de separación del espectro.

[4]La existencia del valor α_0 tendrá sentido más adelante cuando se estudie el error en el procedimiento.

Luego, continuando con el análisis se tiene que bajo H_1 el estadístico R diverge de acuerdo a la observación (iii) dada en la sección 2.3 . Esto es: dado que $k_0 < k$, se tendrá que

$$R \geq \frac{\gamma_k - \gamma_{k+1}}{\gamma_{k+1} - \gamma_{k+2}}$$

Pero γ_{k+1} y γ_{k+2} permanecen acotados dentro del *bulk*, mientras que el valor de γ_k se sale de escala, por lo que R también lo hace. Así, se vuelve altamente probable que R caiga en la región de rechazo del test y por ende no cometa error de tipo II.

La figura 3.6 muestra el comportamiento de la potencia en función del valor de α para los tests de nulas $H_0 : k = 3$ y $H_0 : k = 4$ contra sus respectivas alternativas $H_1 : 3 < k \leq 8$ y $H_1 : 4 < k \leq 8$. Dado que se espera que el valor de β sea muy cercano a 0, la potencia será muy cercana a 1, lo que se evidencia empíricamente en la gráfica. Dicha potencia es calculada usando $N = 15000$ muestras generadas con 8 factores para ambos tests.

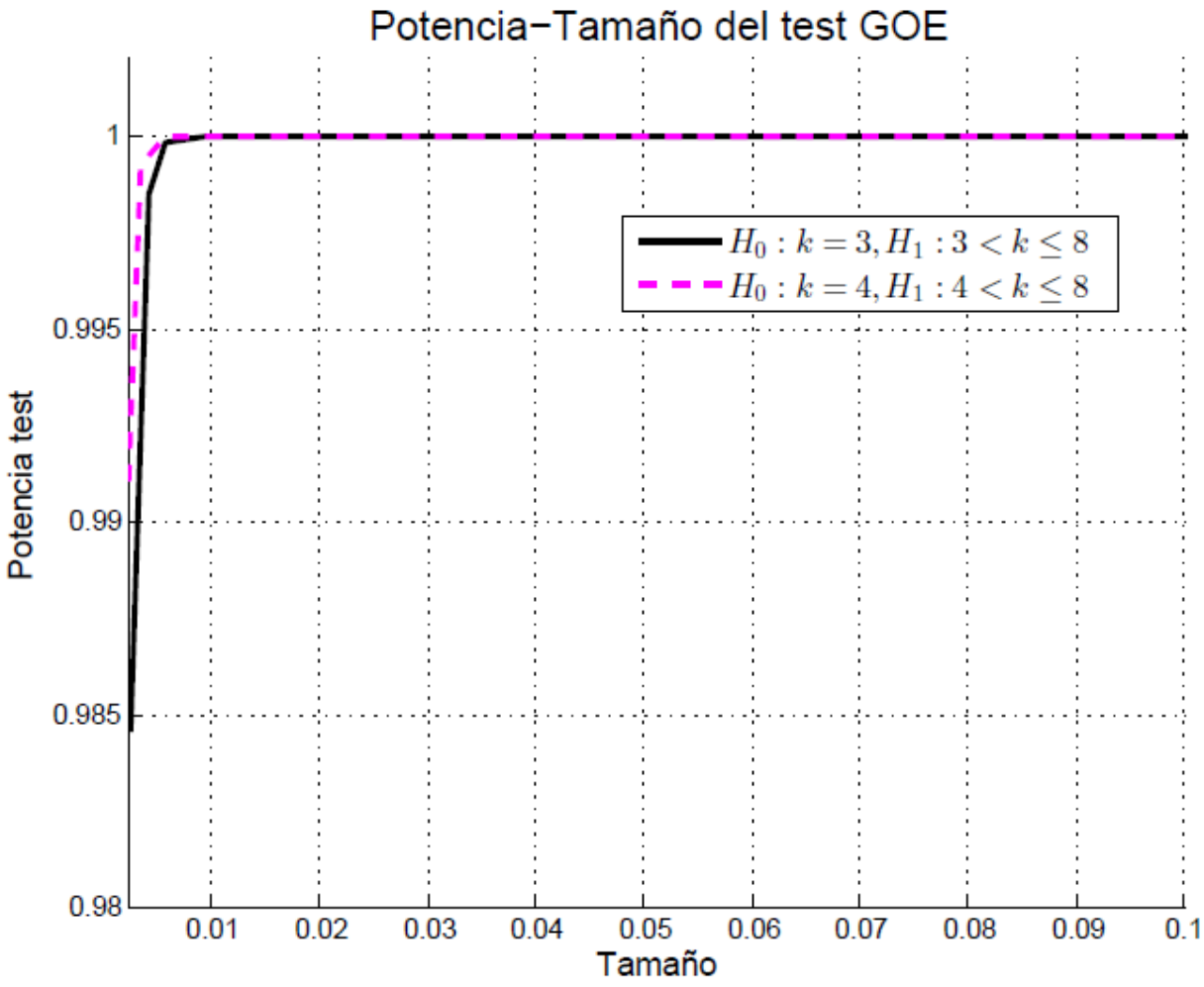


Figura 3.6: Gráfica potencia-tamaño para los tests de nulas 3 y 4 factores.

3.2.3. Estudio de la Gaussianidad

De acuerdo a las evaluaciones anteriores, el test GOE cumple cabalmente con las expectativas que se tienen respecto de su comportamiento.

En particular, se ha puesto en evidencia que el test falla según lo esperado y rechaza una nula cuando es cierta con probabilidad α. Complementariamente, si el test no rechaza la nula cuando es cierta, es posible hablar de un acierto, lo que ocurre, de acuerdo a las pruebas empíricas, con probabilidad $1-\alpha$.

Formalmente se define la probabilidad de acierto del test GOE como:

Definición 3.5 (**Probabilidad de acierto**)

$$\textit{Probabilidad de Acierto} = \mathbb{P}(\textit{No rechazar } H_0 | H_0 \textit{ es cierta})$$

En esta línea, un asunto interesante de evaluar es qué efectos podría tener una alteración de los supuestos del modelo en la probabilidad de acierto del test.

La distribución gaussiana de los términos de idiosincracia es el principal supuesto a poner a prueba. En otras palabras, el propósito es dilucidar qué tan sensible se vuelve el test en materia de acierto cuando se perturba la gaussianidad de la matriz de ruido.

Para lo anterior, se propone el modelo de perturbación siguiente:

$$X = \Lambda F + \mathrm{e} + \varepsilon Z, \tag{3.1}$$

donde Z es una variable aleatoria con distribución no gaussiana y ε es un parámetro ponderador que representa el nivel de pertubación que se quiere alcanzar.

Para ilustrar se presentan tres posibles distribuciones para la variable Z:

(i) Distribución Uniforme.

(ii) Distribución de Laplace.

(iii) Distribución de Pareto.

Las distribuciones de Laplace y de Pareto pueden resultar menos familiares, razón por la cual se definen a continuación en términos de sus densidades.

Definición 3.6 (**Distribución de Laplace**) *Una variable aleatoria tiene una distribución de Laplace de parámetros μ y b si su densidad de probabilidad es:*

$$f(x) = \frac{1}{2b}\mathrm{e}xp\left(-\frac{|x-\mu|}{b}\right) = \begin{cases} \frac{1}{2b}\mathrm{e}xp\left(-\frac{\mu-x}{b}\right) & si\ x < \mu \\ \frac{1}{2b}\mathrm{e}xp\left(-\frac{x-\mu}{b}\right) & si\ x \geq \mu \end{cases}$$

En donde a μ se le conoce como parámetro de ubicación y a $b > 0$ como parámetro de escala. La distribución tiene media μ y varianza $2b^2$.

A modo de observación cabe mencionar que la distribución de Laplace de parámetros $\mu = 0$ y $b = 1$ vista desde el primer cuadrante corresponde a la distribución exponencial escalada por $\frac{1}{2}$. Además, su densidad de probabilidad tiene cierta reminiscencia a la densidad de la distribución normal, salvo que se expresa como diferencia absoluta con respecto a la media, y no como su cuadrado según el caso gaussiano. Como consecuencia, la distribución de Laplace tiene colas más pesadas que la normal. Esto la hace una candidata idónea si de perturbaciones se trata, pues, aunque asumir gaussianidad funciona para modelar variables y datos reales, existen contextos en los que el supuesto de normalidad no basta para obtener resultados plausibles. En tal caso, añadir una perturbación de tipo Laplace podría ser ventajoso debido a que genera un mejor ajuste de los datos reales de la cola.

Por otra parte, la distribución de Pareto aparece en diversos fenómenos que a priori no se relacionan. Ejemplos de esto son: el tamaño de los asentamientos humanos, la distribución del tamaño de archivos en tráfico de Internet[5], el precio estandarizado de retornos de activos, entre otros.

Definición 3.7 (**Distribución de Pareto**) *Una variable aleatoria tiene una distribución de Pareto de parámetros k y σ si su densidad de probabilidad es:*

$$f(x) = k\frac{\sigma^k}{x^{k+1}}$$

En donde a $k > 0$ se le llama parámetro de forma y a σ de escala.

La razón por la que esta distribución es otra buena candidata a perturbación es la misma que para la distribución de Laplace, es decir, incluir una perturbación de tipo Pareto puede ser beneficioso cuando el principal interés es modelar el rango completo de los datos y particularmente de su cola.

En conclusión, ambas distribuciones, de Laplace y de Pareto, son modelos un poco más complejos que, junto a supuestos de normalidad, permiten describir una muestra en su totalidad y aportan mayor realidad al modelamiento.

De acuerdo a lo ya expuesto, el experimento es el siguiente: para un conjunto de $N = 10000$ muestras generadas con 5 factores se calcula la probabilidad empírica de acierto bajo la perturbación del modelo dada por la expresión 3.1. Luego, se repite lo anterior para varios valores de ε, lo que permite evaluar el acierto de acuerdo al nivel o grado de perturbación.

El resultado de tal experimento se muestra en la figura 3.7 , la cual se obtiene a partir de contabilizar cuántas de las N muestras no son rechazadas al aplicárseles el test de nula $H_0 : k = 5$ factores contra $H_1 : 5 < k \leq 10$ para un nivel $\alpha = 0{,}05$. Además, como parte del resultado, se presentan simultáneamente las tres distribuciones estandarizadas para la variable Z, en donde la uniforme tiene soporte en $[-1, 1]$, la de Laplace tiene parámetros $\mu = 0$ y $b = 1$ y la de Pareto, $k = \sigma = \frac{1}{\theta}$ con $\theta = 0{,}3$.

[5]Esto es mediante TCP, protocolo por excelencia usado para cualquier transferencia de archivo en la red.

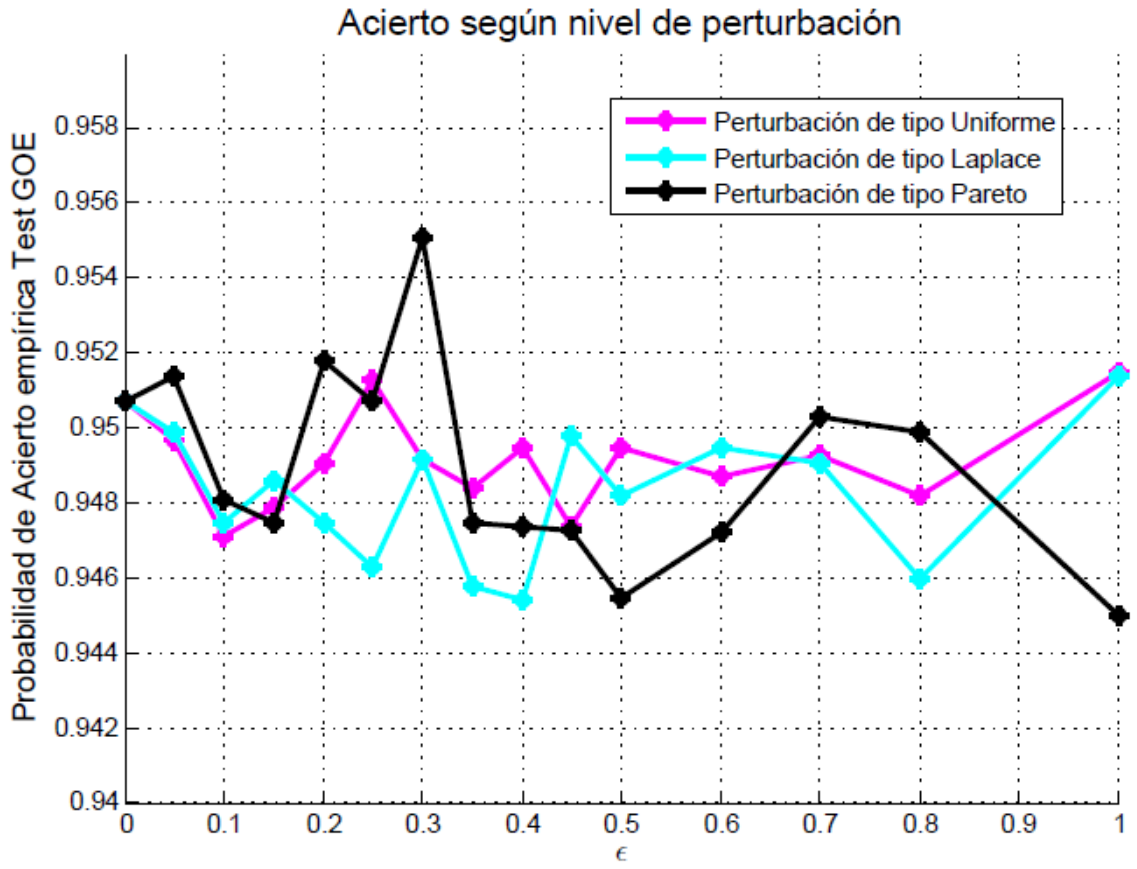


Figura 3.7: Probabilidad empírica de acierto para el test *GOE* según el nivel de perturbación.

Lo que se concluye es que la distribución uniforme es la que genera menos alteraciones del acierto. Por el contrario, la de Pareto parece ser la que más ruido introduce a la gráfica, y por lo tanto, la que perturba en mayor medida la probabilidad de acierto. No obstante, independientemente del tipo de perturbación y del grado de esta, el acierto, en términos generales, se mantiene acorde a lo esperado ($1 - \alpha = 95\,\%$) si se deja a la varianza propia de la experimentación en un segundo plano.

En esta línea, es posible seguir experimentando con el fin de evaluar qué tan sensible es el test a las pertubaciones del modelo, ya sea del supuesto de gaussianidad u otro. Además, pueden definirse otras medidas que evalúen el comportamiento del test. De acuedo al interés del experimentador, el acierto puede no ser el mejor indicador, de modo que si el objetivo es, por ejemplo, minimizar error de tipo II, observar la potencia podría dar más luces del rendimiento del test. En este sentido, la figura 3.12 es un estudio ínfimo respecto del amplio análisis de sensibilidad que puede realizarse.

3.3. Procedimiento de estimación de factores

En esta sección se define formalmente el procedimiento que hace uso del test *GOE* para estimar el número de factores.

Tal como se introdujo en la sección 2.3.1 , el procedimiento consiste en aplicar iterativamente tests *GOE* con hipótesis nulas que van proponiendo distintos números de factores posibles. Así, en el momento en que algún test ya no pueda rechazar su hipótesis nula, el procedimiento se detiene y entrega como número estimado de factores el último que se testeó.

Una vez definido formalmente el procedimiento, se irá dando paso, a lo largo de las secciones siguientes, al estudio de sus características y propiedades, además de evaluar los resultados de las pruebas cuyo principal objetivo es dar a conocer el rendimiento de sus estimaciones.

El procedimiento recibe el nombre de *Procedimiento de Factores* y asume que el verdadero número de factores k no sobrepasa un cierto valor k_1 fijo a priori. Luego, procede como sigue:

(i) Se testea $H_0 : k = 1$ contra $H_1 : 1 < k \leq k_1$ con un nivel de significancia α. Si H_0 no es rechazada, el procedimiento se detiene y el número estimado de factores es 1.

(ii) De lo contrario, si H_0 es rechazada, el procedimiento testea $H_0 : k = 2$ contra $H_1 : 2 < k \leq k_1$ para el mismo valor de α.

(iii) Finalmente, el procedimiento itera los tests hasta que alguna nula H_0 no pueda ser rechazada. Entonces se considera el número correspondiente de factores como el estimado.

Es necesario notar que el procedimiento siempre se detiene, pues realiza a lo más $k_1 - 1$ iteraciones. En este sentido es ventajoso contar con algún margen aproximado de ubicación para el valor k, de manera que sea posible fijar un parámetro k_1 razonable y evitar un número alto de iteraciones y posible sobrestimación.

Aplicando este procedimiento, el número estimado de factores se acercará al verdadero valor con probabilidad $1 - \alpha$ en el regimen $n \to \infty$, lo que se verá reflejado en los resultados mostrados en las secciones venideras.

Observación: Es razonable considerar que un procedimiento descrito de esta forma pueda tener alguna clase de monotonía. Es decir, alguna propiedad que relacione el evento en que rechaza un valor de $k_0 + 1$ con el evento en que rechaza k_0. Intuitivamente podría pensarse que si el procedimiento rechaza $k_0 + 1$, entonces necesariamente rechazó k_0. Esto no ocurre con el procedimiento, es decir, no se tiene una propiedad de monotonía, lo que, en efecto, puede comprobarse considerando lo siguiente:
Pedir una condición de monotonía de este estilo en el procedimiento es equivalente a pedir que

$$\text{Rechazar } k_0 + 1 \Longrightarrow \text{Rechazar } k_0$$

O equivalentemente,

$$R_{(k_0+1)} > \mathcal{V}_{(k_0+1)} \Longrightarrow R_{(k_0)} > \mathcal{V}_{(k_0)}$$

donde $R_{(k_0)}$ y $R_{(k_0+1)}$ son los valores que toma el estadístico R bajo las hipótesis nulas $k = k_0$ y $k = k_0 + 1$ respectivamente, y donde $\mathcal{V}_{(k_0)}$ y $\mathcal{V}_{(k_0+1)}$ son los valores críticos correspondientes a k_0 y $k_0 + 1$ respectivamente.

Pero, R decrece en k_0 por ser definido como un máximo y por su parte, los valores críticos también decrecen en k_0 (para verificarlo basta con observar la tabla 3.3 del anexo A), lo que implicaría que $R_{(k_0+1)} \leq R_{(k_0)}$ y $\mathcal{V}_{(k_0+1)} < \mathcal{V}_{(k_0)}$. Luego, a partir de este par de desigualdades, es imposible concluir algún tipo de orden entre $R_{(k_0)}$ y $\mathcal{V}_{k_0}$, por lo que no se cumple la implicancia mencionada.

3.4. Desarrollo y evaluación del *Procedimiento de Factores*

Toda vez que se ha implementado el procedimiento de factores (ver anexo D implementación 3.6), se describen pruebas empíricas llevadas a cabo como método para evaluar su rendimiento y validarlo en términos de la calidad de las estimaciones que realiza.

Por otro lado, no hay que olvidar que la real utilidad del procedimiento radica en que contar con una buena estimación del número de factores, hará de las predicciones y pronósticos posteriores, procesos más certeros. En pos de lo anterior, para evaluar la calidad de la estimación, es de sumo interés conocer los aciertos y desaciertos del procedimiento.

Definición 3.8 (**Acierto y desacierto**) *Se define como acierto del procedimiento al evento en que $\hat{k} = k$, donde k es el número real de factores y $\hat{k}$ el estimado. Contrariamente, se le llama desacierto al evento en que $\hat{k} \neq k$.*

Como evidencia de que el procedimiento es un buen método de estimación, se muestra más abajo (en la figura 3.8) la tasa o probabilidad empírica de acierto obtenida a partir de $N = 10000$ muestras simuladas con valores que van desde 2 hasta 17 factores.

De la figura 3.8 se desprende que la probabilidad empírica que muestra la línea contínua azul presenta fluctuaciones del orden del 1 %. Porcentaje que se asocia a la aleatoriedad intrínseca de la experimentación. Por su parte, la probabilidad promedio (línea punteada) se ubica muy cercana al 0.95 que corresponde al valor esperado en el regimen asintótico.

Además, dado que la gráfica 3.8 es generada para muestras con distintos números de factores, se observa que el procedimiento se mantiene en general estable con respecto a la dimensión k, lo cual podría considerarse como un buen acercamiento hacia la robustez. Aunque, para hablar de robustez con autoridad, se necesitaría echar a andar muchas otras pruebas del estilo de la que se realizó en la sección 3.2.3 , en la cual se añade una nueva distribución para

perturbar el modelo. Esto es parte de un análisis que este trabajo no desarrolla, de modo que puede ser incluido dentro de una línea de investigación futura.

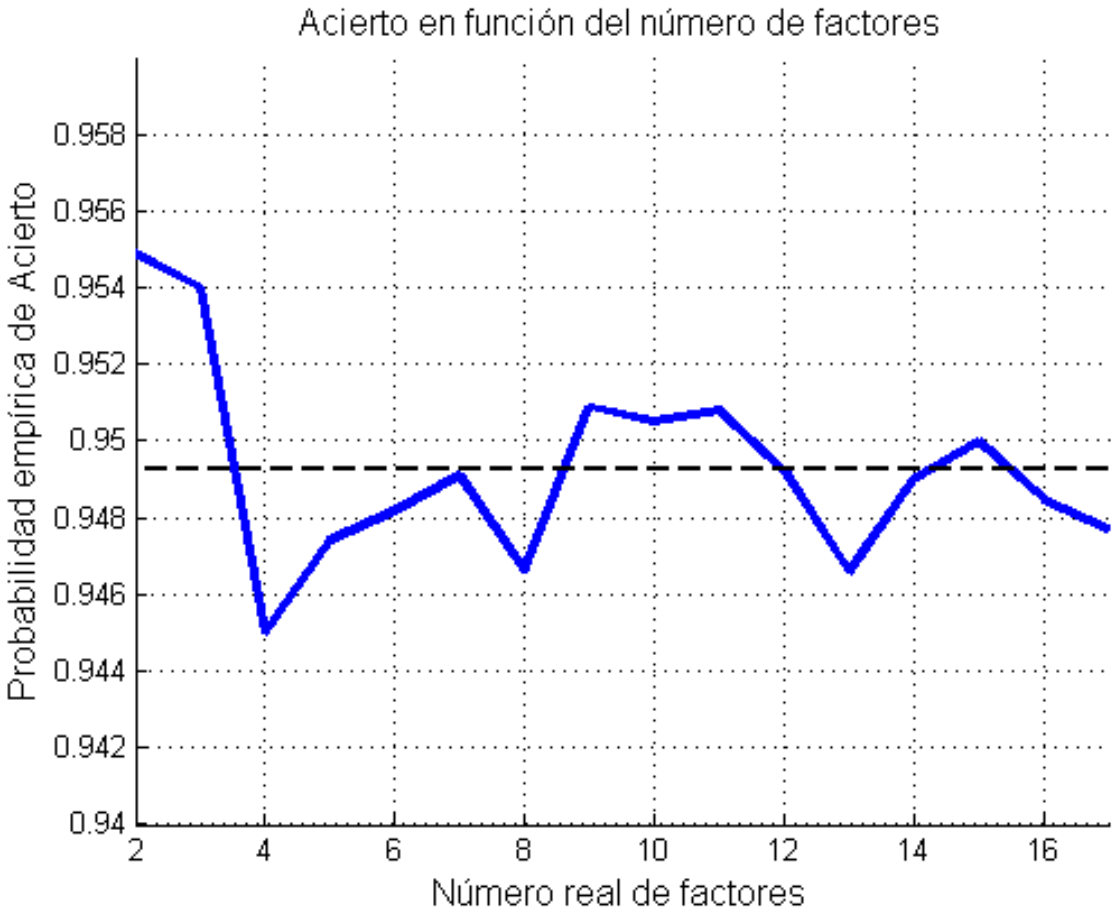


Figura 3.8: Probabilidad empírica de acierto en función del número de factores

Ahora bien, para ratificar que la probabilidad de acierto del procedimiento se acerca al valor esperado $1 - \alpha$ en el regimen asintótico, se muestra en la figura 3.9 la probabilidad empírica obtenida en función de la cantidad N de muestras sobre las cuales se aplica. Allí la varianza asociada a la experimentación va en desmedro, conforme va aumentando la cantidad N de muestras, por lo que es posible apreciar de manera más evidente la convergencia empírica hacia el valor esperado de 0.95.

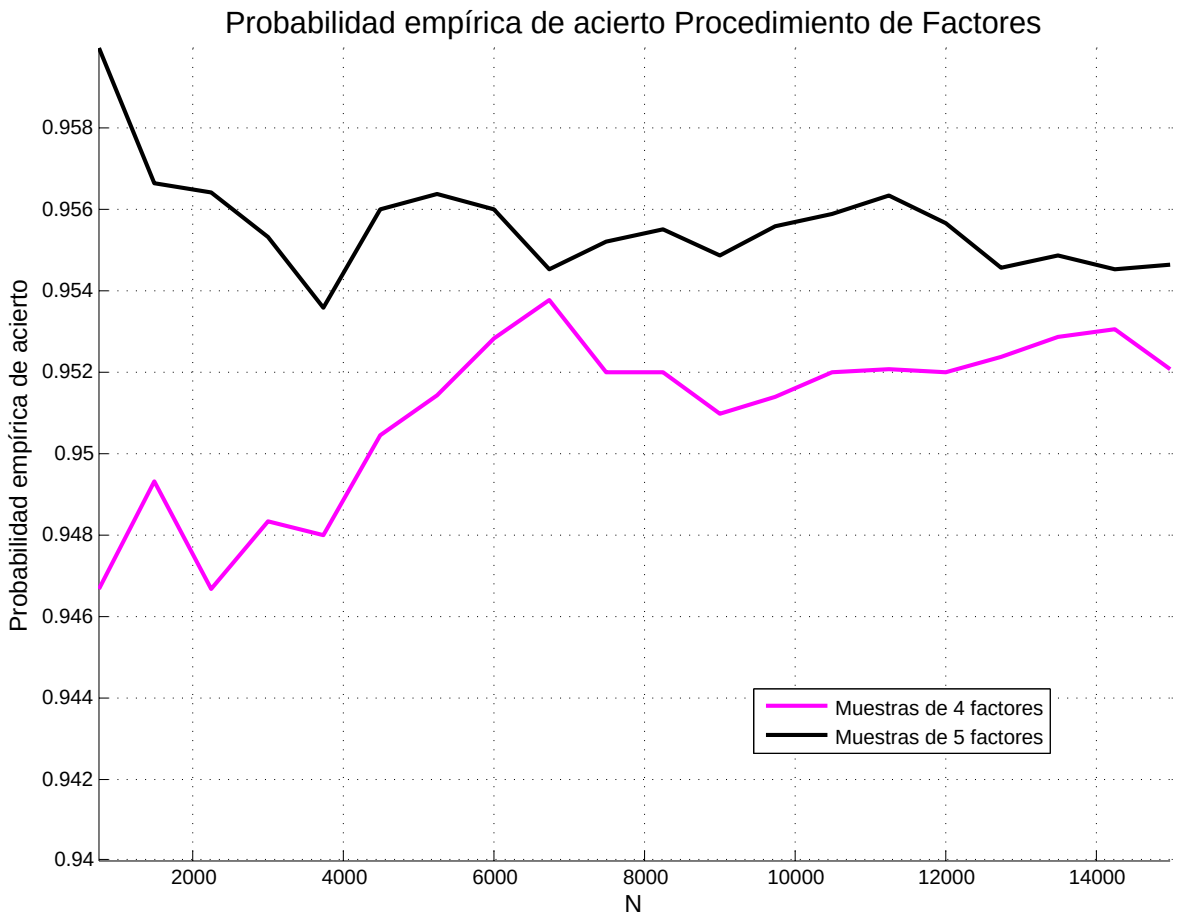


Figura 3.9: Probabilidad empírica de acierto en función del número de muestras.

3.5. Estudio del *Procedimiento de Factores*

Una vez que se ha probado, al menos empíricamente, que el procedimiento funciona como método de estimación del número de factores, es posible estudiar sus características y comportamiento, poniendo en evidencia sus fortalezas y debilidades.

Lo primero a saber es que el evento en que el procedimiento realiza $k_1 - 1$ iteraciones tiene probabilidad no nula. Luego, también tiene probabilidad no nula el evento en que rechaza las $k_1 - 1$ hipótesis nulas, en cuyo caso le es imposible estimar un número de factores. En estas situaciones el procedimiento no aplica y no se obtiene un resultado significativo.

Teóricamente se espera que la intersección de los eventos arriba descritos tenga una probabilidad baja. Mientras que empíricamente, esto se comprueba y, más aún, se evidencia el porqué de la existencia de estos eventos anómalos.

Haciendo un estudio exploratorio de los valores propios de las matrices de covarianzas de muestras en las que el procedimiento no responde, se encuentra un fenómeno común: dos valores propios pertenencientes al *bulk* muy cercanos. Esto significa que la diferencia entre ellos es un valor muy cercano a cero, lo que implica que el denominador del estadístico R disminuye considerablemente. Como consecuencia, R crece sobremanera y todos los tests que realiza el procedimiento son más propensos a rechazar. Este fenómeno de cercanía extrema entre dos valores propios permite explicar las anomalías, pero sin duda está sujeto a la precisión computacional con la que se trabaja, de manera que si se aumenta la precisión, la cantidad de anomalías disminuye.

La tasa con la que ocurren los casos anómalos se muestra en la siguiente tabla.

Tabla 3.1: Tasas casos anómalos

N	Factores en la muestra	Tasa (%)
10000	4	0.73
10000	5	0.84
10000	6	1.15
10000	7	1.64
10000	8	2.28

La tabla 3.1 es construida a partir de N aplicaciones del procedimiento sobre muestras distintas con $3, \ldots, 8$ factores, y con un valor máximo de factores $k_1 = 10$. Del total N de muestras se contabilizan los casos anómalos y se obtiene el porcentaje respectivo.

Si bien la tasa es pequeña, no es nula, por lo que se hace indispensable observar con detalle el origen de estas anomalías. Este es: las diferencias entre aquellos valores propios de la matriz de covarianza respectiva que se encuentran en el *bulk*. Es decir, el estudio se concentra esta vez en $\gamma_i - \gamma_{i+1}$ para aquellos $\gamma_i's$ en el *bulk*.

De acuerdo a lo anterior, se escogen 233 muestras anómalas y se observan las diferencias de los valores propios, amplificadas por $n^{1/6}$ que corresponde al factor de escala acorde a la apreciación macroscópica en el *bulk* para matrices *GOE*. Con esto, se extrae el valor mínimo

de ellas, a partir del cual se busca definir una regla que permita detectar anomalías y evitarlas.

La figura 3.10 muestra la densidad de probabilidad del mínimo de las diferencias amplificadas de acuerdo a la escala del *bulk*. Lo que se busca con ello, es dar una pauta de exclusión de los casos anómalos.

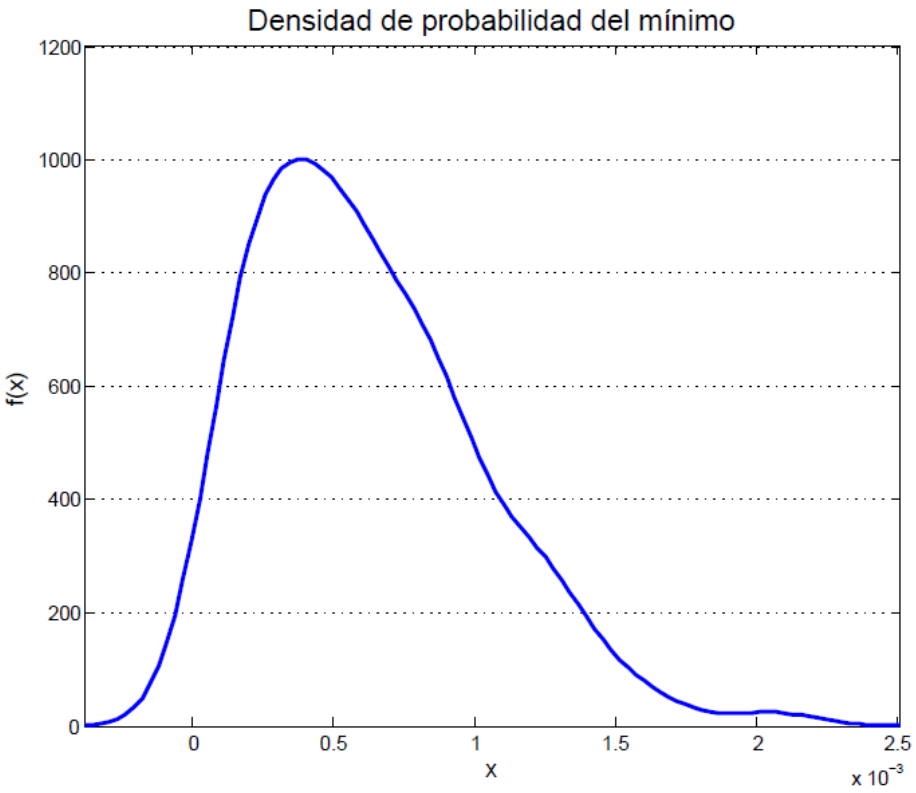


Figura 3.10: Densidad de probabilidad para el mínimo de las diferencias de valores propios.

Dicha pauta viene dada por una cota superior de las diferencias entre valores propios. Se considera el percentil 80 de la población del valor mínimo, que se condice con establecer un umbral en $9 \cdot 10^{-4}$ para la figura 3.10. Esta medida constituye un criterio heurístico que resulta razonable dado el contexto y la precisión computacional[6] con la que se trabaja.

Así, la pauta a seguir se enuncia como sigue:
Sean $\gamma_1 \geq \cdots \geq \gamma_n$ los valores propios de la matriz de covarianza de la muestra. Para aquellos casos en que $\gamma_i - \gamma_{i+1} < \eta \cdot n^{-1/6}$ para algún i y para $\eta = 9 \cdot 10^{-4}$, el Procedimiento de Factores no es significativo.

Observación: Claro está que la pauta aquí presentada no es una regla categórica, por lo que el experimentador es libre de fijar el umbral de acuerdo a sus intereses y a las pruebas que desee realizar.

[6]Precisión computacional de 4 dígitos decimales

Continuando con el estudio del procedimiento, otra de sus características a mencionar, y que viene heredada del test *GOE*, tiene relación con la coherencia en la información obtenida a partir de su aplicación.

Ya es sabido que cuando el procedimiento no puede rechazar la nula de $H_0 : k = k_0$ se detiene y estima k_0 factores. Ahora bien, considerando un escenario hipotético en donde se le permita al procedimiento avanzar una iteración más para testear la nula de $k = k_0 + 1$, lo natural sería que no la rechazara. Siguiendo este razonamiento tampoco debiese rechazar $k_0 + 2, k_0 + 3, \ldots$ factores cuando avanza $2, 3, \ldots$ iteraciones más.

Lo anterior es equivalente a preguntarse qué resultado arrojaría el test *GOE* de nula $H_0 : k = k_0$ contra $H_1 : k_0 < k \leq k_1$ si se aplica a muestras simuladas con $k_0 - 1$ factores. Medir la ocurrencia de estas situaciones, aunque carezcan de sentido, podría ser útil en el contexto del procedimiento si se considera como una de sus debilidades.

Luego, si el procedimiento no rechazara la hipótesis nula de $H_0 : k_0 + 1$ bajo el escenario hipótetico anteriormente descrito, entonces esto indicaría que existe cierta lógica por parte del procedimiento. Sin embargo, debido a la carencia de una propiedad de monotonía en el sentido de la mencionada en la observación 3.3 , existe una probabilidad no nula de que el procedimiento rechace y pierda coherencia en este contexto.

A modo de experimento y con el fin de cuantificar dicha pérdida, se aplica el procedimiento sobre $N = 15000$ muestras con 5 factores, y como resultado arroja que el $94.7\,\%$ de las veces no es posible rechazar la existencia de 5 factores. Porcentaje que se enmarca dentro de la probabilidad empírica de acierto ya probada para $\alpha = 5\,\%$, y dentro del cual, un $0.37\,\%$ aproximado de las veces, el procedimiento rechaza la nula de 6 factores perdiendo coherencia. Este ejercicio se repite para diversas muestras simuladas con diferentes números de factores, obteniéndose porcentajes muy similares, lo cual permite establecer que la tasa promedio con la que se pierde coherencia fluctúa alrededor del $10\,\%$ del valor de α (ver tabla 3.5 en el Anexo E). Por tanto, resulta razonable despreciarla para efectos de pruebas y estudios empíricos.

3.5.1. Sobrestimación y subestimación

Otro tema de estudio tiene relación con el desempeño del procedimiento en términos de los errores que comete en la estimación.

Es claro que dentro del $5\,\%$ aproximado de error que comete el procedimiento existe tanto sobrestimación como subestimación del número de factores. Sin embargo, el procedimiento de factores se ha caracterizado por no seguir las expectativas convencionales en materia de tests de hipótesis. Así es como, al realizar estudios empíricos, se observa que la sobrestimación y subestimación presentan asimetrías.

Este fenómeno surge a raíz de la asimetría que existe entre las tasas empíricas de error de tipo I y II. En particular, haciendo la analogía, es posible asociar la sobrestimación con α y la subestimación con β.

Para comprender la relación entre α y la sobrestimación, debe entenderse esta última como el evento en que el procedimiento realiza más iteraciones de las que debería. Puesto de otra forma, esto significa que cuando el procedimiento testea el número verdadero de factores lo rechaza, lo que se asocia al concepto de error de tipo I. Luego, dado que el error empírico de tipo I se comporta de acuerdo a lo esperado según el tamaño nominal utilizado, es de suponer que la tasa de sobrestimación también lo haga.

Por otra parte, para entender la correspondencia entre β y la subestimación, debe considerarse que esta última equivale a conseguir que el procedimiento de factores se detenga antes de llegar a testear el número verdadero k de factores. En otras palabras, el procedimiento realiza menos iteraciones de las que debe y no rechaza la nula de $H_0 : k = k_0$ cuando $k_0 < k$, lo que equivale al concepto de error de tipo II puesto que la hipótesis alternativa es la corresponde a la verdad.

Otro punto importante de destacar en relación a la sobre y subestimación, es el carácter unilateral del test. Es decir, el hecho de que la hipótesis alternativa del test se defina como $H_1 : k_0 < k \leq k_1$ y no como $H_1 : k \neq k_0$, podría estar incidiendo en la asimetría que existe entre sobre y subestimación.

Esta última puede apreciarse, por ejemplo, para el caso de un tamaño nominal $\alpha = 5\,\%$, y para N muestras simuladas con $3, 4, \ldots, 8$ factores a las que se les aplica el procedimiento con un número máximo de factores $k_1 = 10$. Como resultado se obtiene la tabla 3.2 , en la cual se observa que la subestimación es nula, mientras que la sobrestimación tiene tasa positiva y con tendencia a la baja según la chance menor que se tiene de sobrestimar cuando el número de factores de la muestra va en aumento.

Tabla 3.2: Tasas de sobretimación y subestimación

N	Factores en la muestra	Tasa Subestimación(%)	Tasa Sobrestimación(%)
30000	3	0	4.4367
10000	4	0	4.5400
10000	5	0	4.0900
10000	6	0	4.0700
10000	7	0	3.9100
10000	8	0	2.9700

Observación: A modo de observación cabe mencionar que si a la tasa de casos anómalos se le suma la de sobre y subestimación de la tabla 3.2 , se obtiene un porcentaje aproximado de $5\,\%$ que corresponde al error. Puesto de otra forma, según la definición de desacierto dada por 3.8 , la tasa de anomalías forma parte del error, por lo que esta se compensa con la de la sobrestimación para mantener el nivel de error en $5\,\%$.

Luego, solo para ilustrar una referencia acerca de la sobrestimación en el número de factores, se muestra en la figura 3.11 el histograma de la cantidad de veces en que el estimado sobrepasa el número real de factores. El histograma contempla un total de $N = 10000$ aplicaciones del procedimiento sobre muestras generadas con 4, 5 y 6 factores para un parámetro $k_1 = 10$.

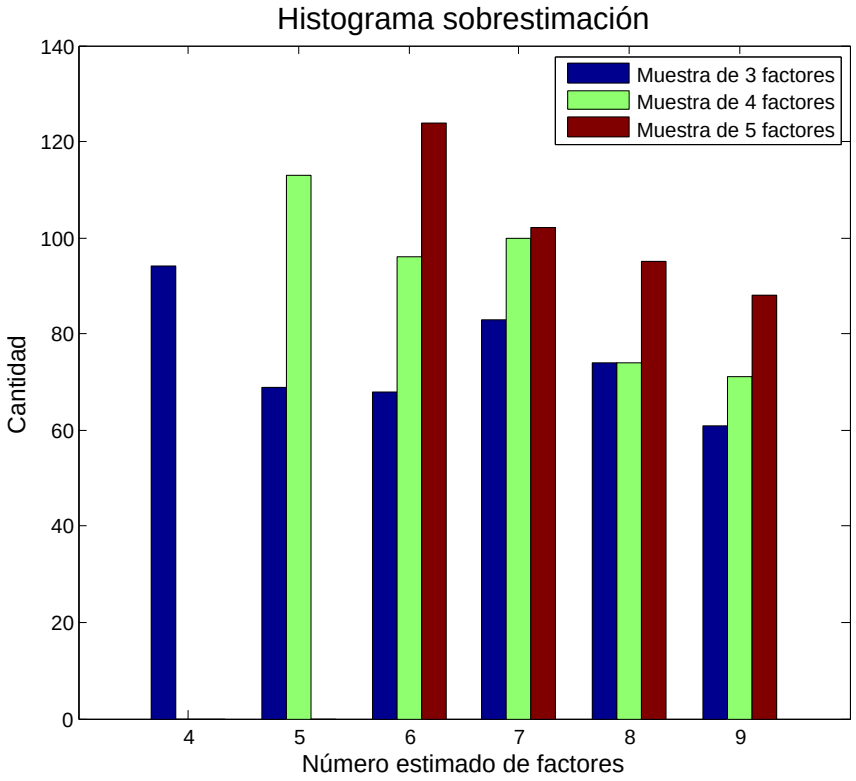


Figura 3.11: Histograma de sobrestimación para muestras de 3, 4 y 5 factores

Finalmente, dentro de este análisis no hay que olvidar que las tasas de sobre y subestimación anteriormente expuestas, son obtenidas a partir de un parámetro $\alpha = 5\,\%$. Por lo que, en particular, para el caso de la tasa de subestimación, podría tenerse un escenario distinto si se utilizan otros rangos de α. Este supuesto es inducido por la gráfica 3.6, en donde se tiene que, para valores pequeños de α ($\alpha << 0.01$), existe un error empírico de tipo II no nulo en el test *GOE*, el cual podría repercutir en el procedimiento alterando las tasas, y por ende la proporción entre subestimación y sobrestimación.

3.5.2. Estudio paramétrico del *Procedimiento de Factores*

Cuando ya se ha analizado el error en la estimación por parte del procedimiento, es posible realizar un análisis de los parámetros que utiliza. A saber: el tamaño nominal α y el número máximo de factores k_1.

El conocimiento con el que se cuenta hasta el momento sobre este último, es que si se fija con una holgura razonable con respecto al número real de factores, entonces la probabilidad de acierto se mantiene dentro del margen establecido por $1-\alpha$. De lo contrario, fijar un valor muy grande para el parámetro k_1, hará que el número estimado de factores se sobrestime en mayor medida.

Luego, cabe preguntarse cuál sería una holgura adecuada para el k_1 y qué restricciones tiene su valor si se quiere conseguir un nivel de acierto determinado. Por tanto, una vez más se hace necesaria la experimentación para dar respuesta a estas y otras preguntas que puedan surgir en el proceso.

Dado que la principal aspiración es mantener la probabilidad de acierto acorde a $(1-\alpha)$, se realiza el siguiente estudio empírico:

En primer lugar, para observar la real influencia que tiene el parámetro k_1 sobre la probabilidad de acierto, lo más conveniente es establecer el valor que se quiere alcanzar de esta última, lo que implica fijar el valor de α, por ejemplo en $5\,\%$. Luego, sobre un conjunto de $N = 10000$ muestras que se generan para 3, 4 y 5 factores, se aplica el procedimiento variando el valor de k_1 que va desde 4, 5 y 6, respectivamente, hasta 18 factores.

Finalmente, se procede a contabilizar el acierto y obtener su probabilidad empírica, la cual se muestra en la figura 3.12. En ella, a pesar del ruido propio del carácter experimental del estudio, no es posible observar ninguna tendencia o influencia clara por parte del k_1 sobre el acierto del procedimiento.

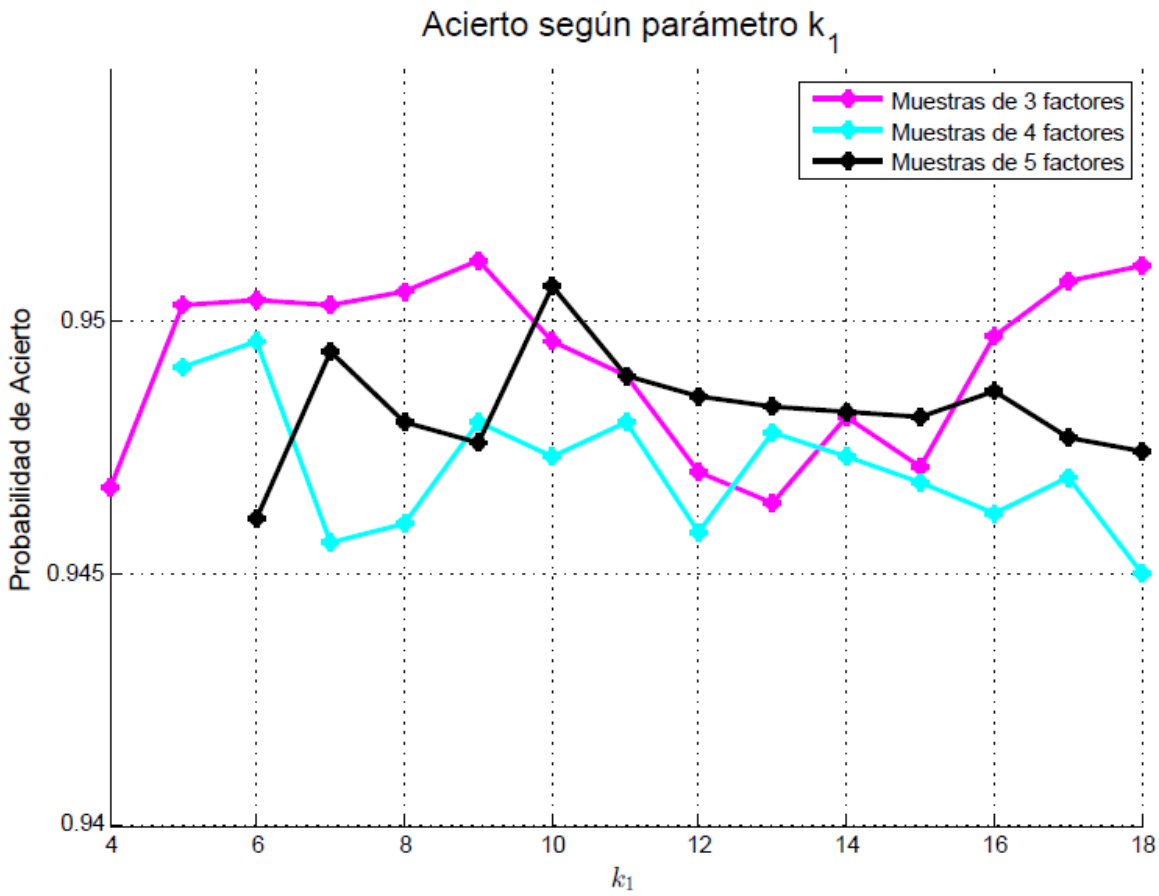


Figura 3.12: Probabilidad empírica de acierto en función de k_1.

De acuerdo a lo anterior, no es posible obtener ninguna conclusión más allá que afirmar que el valor del parámetro k_1 no tiene mayor efecto en el acierto, debido, probablemente, a la falta de una condición de monotonía en el procedimiento. Es decir, la probabilidad de acierto no se ve afectada por el parámetro k_1 porque no se puede establecer un orden de eventos que involucren rechazos de valores consecutivos de k_1. Esto sucede por la misma razón que la mencionada en la observación 3.3 , salvo que esta vez, el estadístico R y los valores críticos crecen con k_1.

No obstante, se hace énfasis en el hecho de que la prueba ha sido efectuada para $\alpha = 5\,\%$ y para muestras con unos pocos factores (4, 5 y 6), por lo que se hace necesario continuar con la experimentación, especialmente con aquella centrada en α.

Para ello, lo que se busca es una idea gráfica de cómo se comporta la probabilidad de acierto con cambios en el parámetro α. En particular, se efectúa el ejercicio de aplicar el procedimiento a muestras de $k = 2, 3, \ldots, 17$ factores para cuantificar el acierto en función de k y para distintos valores α. El procedimiento es aplicado con un número máximo de factores de $k_1 = 18$ y el resultado se presenta a continuación.

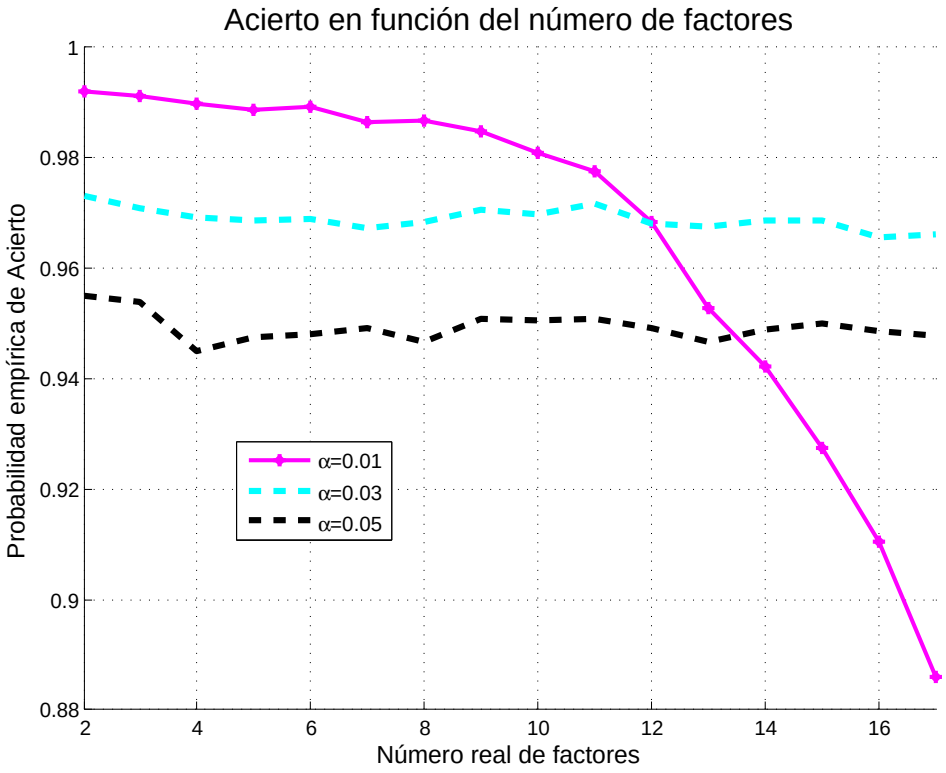


Figura 3.13: Probabilidad empírica de acierto según el número real de factores.

Lo que muestra la figura 3.13 es realmente interesante puesto que se observa cómo para un nivel de α =0.01, la probabilidad decae a medida que el procedimiento se aplica para muestras con un número mayor de factores. En particular, para muestras de más de 8 factores es clara la tendencia a la baja, y se acentúa para muestras de más de 14 factores, en donde la probabilidad se encuentra por debajo del 95 %.

Mediante una indagación empírica exploratoria que se realiza de este nuevo fenómeno, se descubre que para $\alpha = 1\,\%$ y muestras simuladas con un número grande de factores, el procedimiento subestima y entrega como resultado la existencia de 1 factor. En otras palabras, la exigencia del α es tal que el procedimiento se detiene después de haber efectuado una iteración. Lo anterior tiene sentido si se considera que un descenso en el parámetro α significa un aumento en los valores críticos en los que se basa el procedimiento para decidir. Lo cual trae como consecuencia una reducción de la zona de rechazo, y por ende, un aumento de la zona de aceptación. Esto implica que el procedimiento es más propenso a aceptar, por lo que tenderá a detenerse luego de efectuar la primera iteración.

La comprensión de este fenómeno también permite explicar por qué no es factible considerar un parámetro $\alpha \to 0$, pues un error de tipo I tendiendo a 0, hará que la tasa de subestimación se dispare, lo que implica que el error tenderá al 100 % debido a que el procedimiento se detiene tras primera iteración.

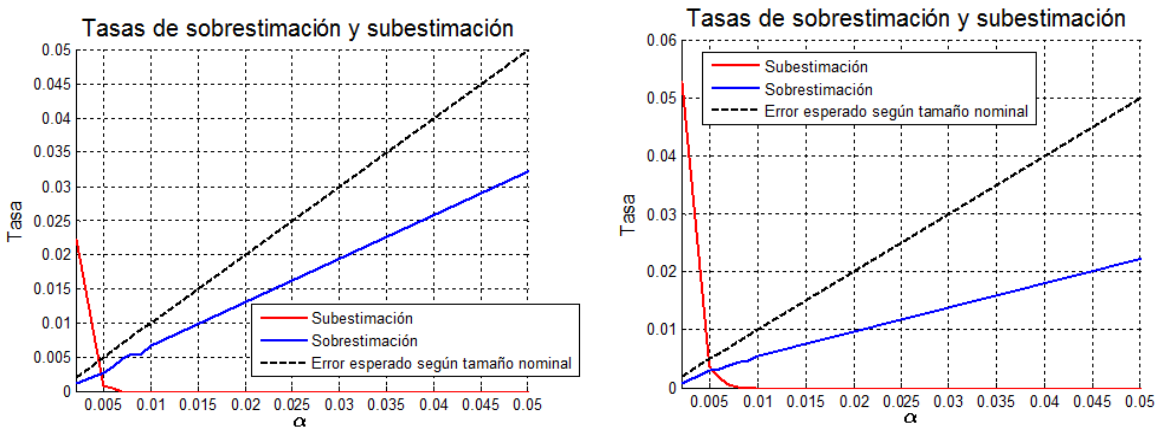
Esto sugiere la existencia de un *trade-off* para el parámetro α. Es decir, una disminución considerable del error de tipo I invalida al procedimiento como método de estimación pues este arroja constantemente como resultado un factor. Por el contrario, si se aumenta sobremanera el valor de α se tendrá mayor error de tipo I, pero se garantizará que el procedimiento realice más de una iteración y que, por ende, tenga la chance de testear el verdadero número de factores.

3.5.3. *Trade-off* para α

De acuerdo a los resultados anteriores, es posible hablar de la existencia de un *trade-off* entre error de tipo I y subestimación. Entendiéndose esta última como el evento en que el procedimiento se detiene tras ejecutar la primera iteración.

Específicamente, el *trade-off* consiste en que, a menor α existe una alta probabilidad de subestimar, por lo que el procedimiento pierde efecto y arroja un factor, lo que conlleva un incremento en la tasa de subestimación. Por otra parte, a mayor α, el problema de la subestimación desaparece, pero se eleva el nivel de error.

Para medir la ganancia o pérdida de las variables involucradas en el *trade-off*, se muestra en la figura 3.14 las tasas de sobre y subestimación para $N = 10000$ muestras de 5 y 7 factores. Para ambos conjuntos de muestras se aplica el procedimiento con un número máximo de factores $k_1 = 10$.



(a) Muestras de 5 factores

(b) Muestras de 7 factores

Figura 3.14: Tasas de sobre y subestimación

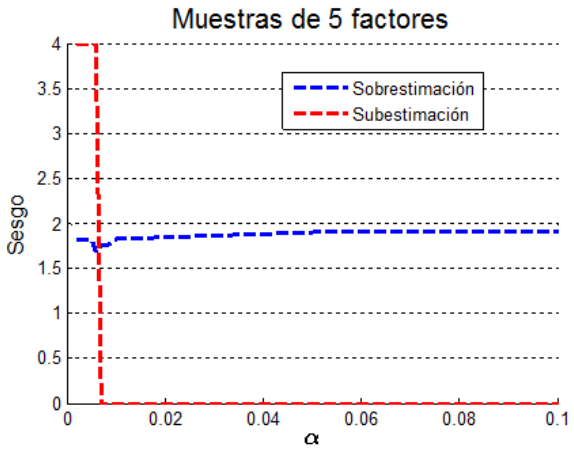
A partir de la figura 3.14 se pueden fijar dos ideas que se habían gestado en esta sección. La primera de ellas es que la figura permite visualizar de forma gráfica la asimetría entre las tasas de sobre y subestimación independientemente de cuál sea el valor que tome α. Por otra parte, la segunda es que se evidencia cómo la subestimación se dispara muy por encima del error esperado para valores pequeños de α, tal como lo sugiere el *trade-off*.

Ahora bien, una vez que se han calculado las tasas respectivas de sobre y subestimación, cabe preguntarse en cuánto se subestima o sobrestima el número de factores. Una ligera noción de esto último la daba el histograma 3.11 para ciertas muestras, sin embargo, se puede realizar otro estudio más detallado como el que se describe a continuación:

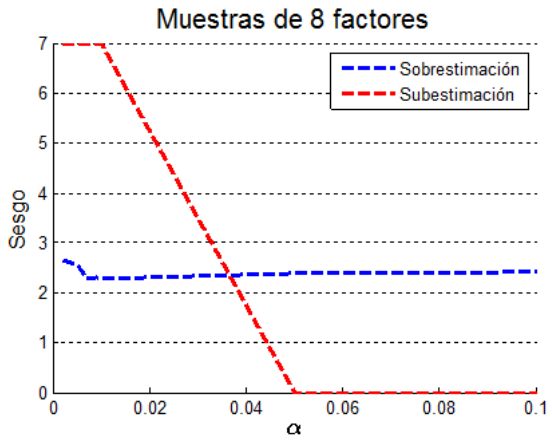
Para calcular cuánto sobre y subestima el procedimiento, se debe estudiar el sesgo, es decir, la diferencia $\mathbb{E}(\hat{k}) - k$, cuya parte negativa equivale a la subestimación, y la positiva, a la sobrestimación.

Luego, para lograr el cálculo empírico del sesgo, se promedian los números estimados de factores que resultan de $N = 10000$ aplicaciones del procedimiento. Esto se efectúa para varios valores de α, y se grafica por separado la parte negativa y la positiva, las cuales en la figura 3.15 aparecen como curvas de subestimación y sobrestimación respectivamente.

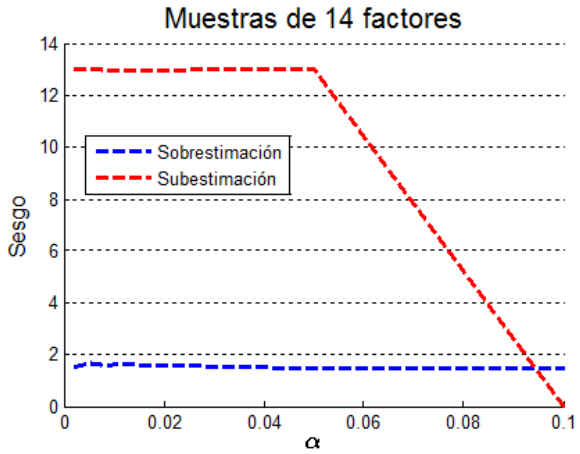
La figura 3.15 muestra el cálculo empírico del sesgo para tres conjuntos de muestras: $N = 10000$ muestras simuladas con 5 factores; otras $N = 10000$, simuladas con 8; y las últimas $N = 10000$, generadas con 14 factores. En los tres casos, se evidencia que la sobrestimación permanece, en general, constante, mientras que con la subestimación ocurre un fenómeno distinto. La curva que la representa se mantiene constante a lo largo de un tramo de α para luego decaer linealmente a 0.



(a) Muestras de 5 factores



(b) Muestras de 8 factores



(c) Muestras de 14 factores

Figura 3.15: Sobre y subestimación para muestras de 5, 8 y 14 factores

Además, si se denota como α_c al último punto de los $\alpha's$ en donde la subestimación permanece constante y como α_0 al punto en el que cae a 0, lo que resulta curioso del resultado mostrado en la figura 3.15 es que dichos puntos dependen del número k de factores presentes en la muestra. Esto podría interpretarse como una especie de transición de fase, y de ser así, su visualización se obtendría a partir de graficar los puntos α_c y α_0 en función de k.

La gráfica 3.16 muestra las curvas $\alpha_c(k)$ y $\alpha_0(k)$ que representan la transición de fase. Los puntos bajo la curva $\alpha_c(k)$ conforman la *zona de subestimación alta*, es decir, el rango para los valores de α en los que se corre el alto riesgo de subestimar el número de factores por 1. Por otra parte, los puntos sobre la curva $\alpha_0(k)$ corresponden a la *zona de subestimación nula*, en otras palabras, equivale a los valores de α para los cuales se garantiza que el procedimiento no subestimará. Finalmente, los puntos entre ambas curvas se pertenecen a la *zona intermedia*, en la cual el efecto es una amalgama de las otras dos zonas que se condice con el decaimiento lineal que se evidenciaba en la figura 3.15 .

Luego, en el contexto de aplicar el procedimiento a muestras para las que se tiene alguna idea aproximada del número real de factores, la figura 3.15 resulta bastante útil. Esto pues, dado un número aproximado de factores $\tilde{k}$, la transición de fase indica el rango del α acorde a los intereses. Por ejemplo, si la prioridad es eliminar la subestimación convendrá fijar un valor de α dentro de la zona de subestimación nula, es decir, un valor tal que $\alpha \geq \alpha_0(\tilde{k})$. Por el contrario, si se permite en algún grado la existencia de subestimación, entonces podría tomarse un α en la zona intermedia, o sea, un valor que satisfaga $\alpha_0(\tilde{k}) < \alpha < \alpha_c(\tilde{k})$. Finalmente, un valor de $\alpha < \alpha_c(\tilde{k})$ perteneciente a la zona de subestimación alta no es recomendable pues el procedimiento pierde efecto.

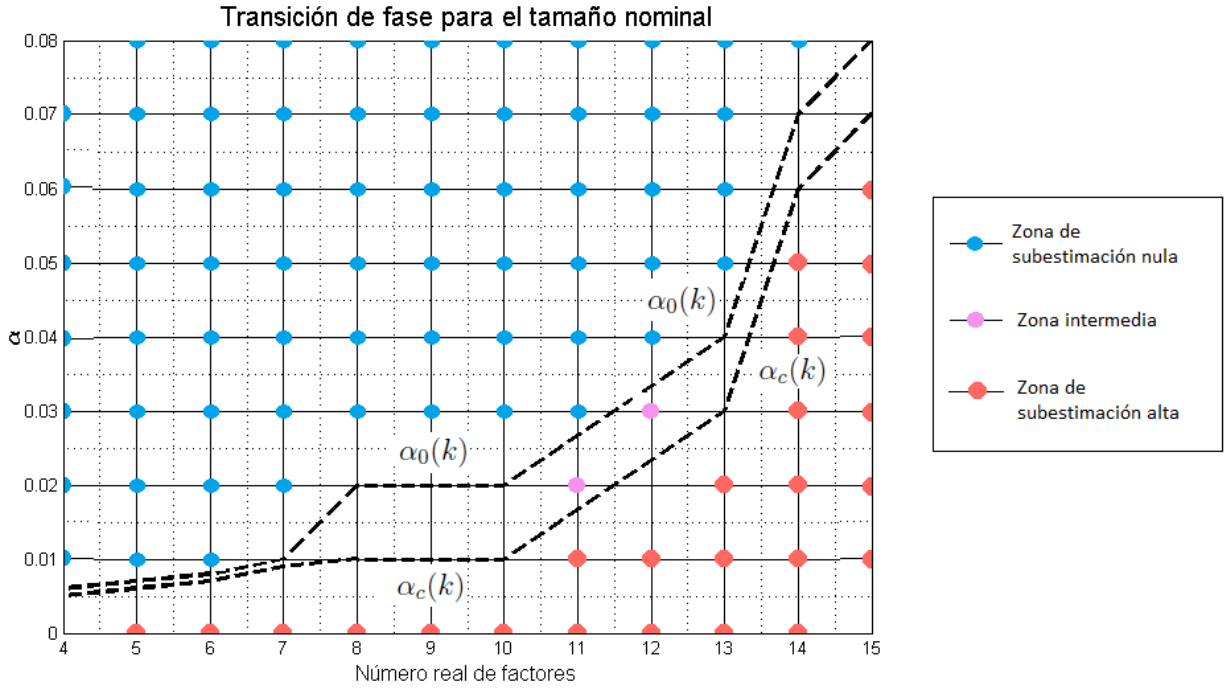


Figura 3.16: Transición de fase para el parámetro α. La transición de fase se ha obtenido para los puntos marcados de la grilla.

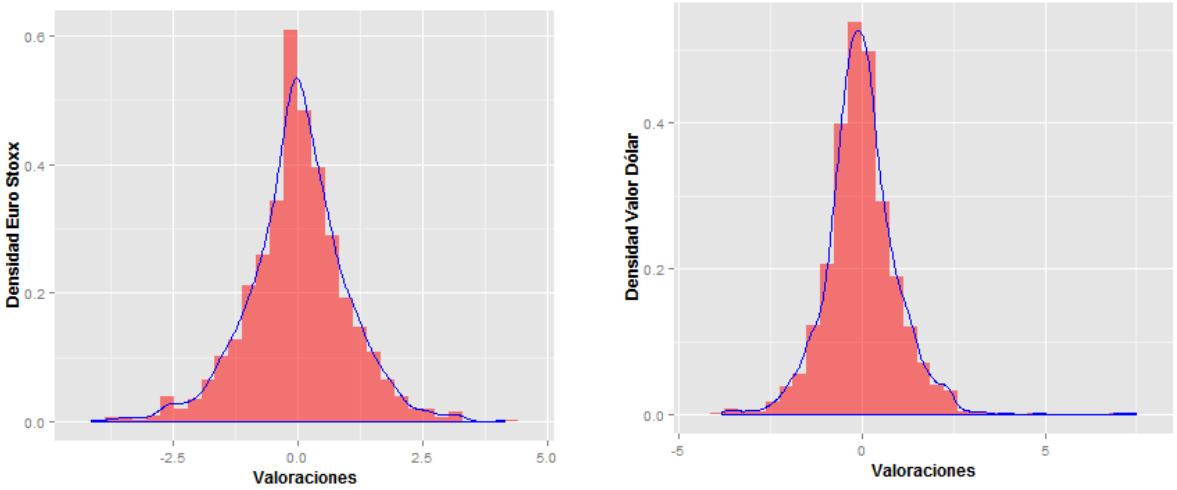
3.6. Aplicación

A modo de cierre del capítulo y del trabajo realizado, se presenta una aplicación del *Procedimiento de Factores* a datos financieros reales otorgados por el Laboratorio de Simulación Estocástica y Estadística del Centro de Modelamiento Matemático (CMM) de la Universidad de Chile.

En esta sección se presentan los datos, se describe en qué forma se aplica el procedimiento sobre ellos, y se comenta el resultado. A modo de anticipo, el resultado obtenido de la aplicación no es significativo, pero de todas formas se deja registro de él y se discuten las posibles causas de la falla. Esto último con el fin de brindar antecedentes al experimentador que desee continuar una línea de investigación aplicada del *Procedimiento de Factores*.

Con respecto a los datos, estos corresponden a series de tiempo de 20 variables que comprenden el periodo que va desde el 17 de Septiembre del 2010 hasta el 21 de Enero del 2015, contando así con 1128 observaciones. Respecto de las variables, se trata de indicadores y activos financieros locales e internacionales. Dentro de ellas, es posible identificar cuatro grandes grupos: bolsas internacionales más influyentes en la moneda chilena, monedas internacionales, principales *commodities* como el petróleo y el cobre y, por último, variadas tasas de interés.

En la Figura 3.17 se muestran histogramas de dos de las series de tiempo estandarizadas. A la izquierda se aprecia el *Euro Stoxx*, índice de la bolsa de Londres, mientras que a la derecha la variación diaria del valor del dólar con respecto al peso chileno (CLP).



(a) Serie de tiempo *Euro Stoxx*

(b) Serie de tiempo dólar en CLP

Figura 3.17: Histogramas y densidades de dos series de tiempo.

El objetivo de esta sección es presentar una aplicación que consiste en hacer uso del *Procedimiento de Factores* sobre una ventana temporal de los datos, lo cual debe entenderse como la aplicación del procedimiento a una porción de observaciones en el tiempo. Como consecuencia de ello, se obtiene un número estimado de factores que depende de la ventana considerada. Lo que sigue, naturalmente, es aplicar el procedimiento a la misma ventana

pero desplazada termporalmente, de manera de obtener, luego de todo el proceso, el número estimado de factores en función de la ventana móvil de tiempo.

Para los datos financieros considerados, se tiene que $n = 20$ y $T = 1128$. Además, para llevar a cabo la aplicación, debe determinarse el valor de los siguientes parámetros:

(i) k_1: Número máximo de factores.

(ii) α: Tamaño nominal con el que se implementa el procedimiento.

(iii) Δ_t: Largo de la ventana temporal a considerar, con $\Delta_t << T$.

(iv) t: Variable que corresponde a las observaciones en el tiempo, así $t = 1, \ldots, T - \Delta_t$.

El detalle de la modalidad en que se aplica el procedimiento sigue a continuación:

Sea una ventana de tiempo de largo Δ_t con inicio en la observación $t = 1$, a la cual se le aplica el procedimiento con parámetros α y k_1 determinados. Lo que se obtiene es un número estimado de factores $\hat{k}(t = 1)$. Luego, si la ventana se inicia en la observación $t = 2$, se obtiene $\hat{k}(t = 2)$. Así, sucesivamente, al ir desplazando el inicio de la ventana a lo largo de la variable $t = 1, \ldots, T - \Delta_t$, se genera la función $\hat{k}(t)$.

El propósito de las ventanas móviles y de la obtención de la función $\hat{k}(t)$ consiste en detectar en qué momento se produce un cambio en el número de factores de la datos financieros. Interesa, particularmente, visualizar dichos cambios para así evaluar, dado que se cuenta con datos históricos, si se condicen con fenómenos económicos reales.

Sin embargo, tras varios intentos por aplicar el procedimiento con la modalidad de ventanas móviles, no fue posible obtener una visualización razonable de la función $\hat{k}(t)$, pues esta presentaba formas distintas cada vez que se modificaban los parámetros. Solo a modo de antecedente, para parámetros $\alpha = 10\,\%$, $k_1 = 6$ y $\Delta_t = \frac{T}{2}$, se observaron saltos en la función $\hat{k}(t)$ desde 2 a 1 factor, y luego de 1 a 3 factores, además se obtuvo como resultado 3 factores luego de aplicar el procedimiento sobre la matriz de datos de dimensiones $n = 20$ y $T = 1128$.

Dentro de las posibles causas que imposibilitan la obtención de resultados para el caso de los datos descritos, se pueden nombrar dos:

(i) La primera tiene relación con el desapego de los datos con el modelo de factores, es decir, es altamente probable que los datos, específicamente, las variaciones de las variables económicas que los componen, no cumplan los supuestos de independencia asumidos en el modelo. Analizando esto desde el punto de vista teórico, se concluye que levantar el supuesto de independencia implica que la matriz de covarianza de los datos se degenera, lo que aumenta la probabilidad de tener valores propios muy cercanos (si no iguales). En consecuencia, no se observa el efecto de repulsión entre ellos y por tanto, la teoría no aplica.

(ii) Por otro lado, la segunda posible causa tiene que ver con el pre-tratamiento de los datos. En general este puede consistir en: tomar diferencias de primer o segundo orden de las

variables o del logaritmo de ellas, o bien, una mezcla de ambas. Además, tal tratamiento dependerá fuertemente de la clase de indicador económico al que se le aplique. Luego, dada la amplia naturaleza económica de las variables financieras involucradas, estas requerirían un pre-tratamiento diferenciado, el cual solo podría brindarse si se cuenta con el conocimiento experto en materias de economía.

A modo de cierre, cabe mencionar que esta modalidad de aplicación del procedimiento de factores tiene cabida dentro del trabajo futuro de investigación, pues si se perfecciona, podría constituirse como una herramienta útil para la detección de cambios o alteraciones en la economía.

Conclusiones y Trabajo Futuro

A modo de conclusión general se establece que se alcanzan los objetivos planteados desde un comienzo, referidos a estudiar el test en su versión *GOE* y a luego darle utilidad por medio del desarrollo del *Procedimiento de Factores*.

En particular, con respecto a la primera de las metas planteadas en el resumen, las pruebas presentadas en la segunda sección del capítulo 3 y los resultados obtenidos de ellas, permiten ampliar el conocimiento que se tiene del test en su versión *GOE* y concluir que no posee ninguna diferencia respecto de la versión *GUE*. En consecuencia, ambos son válidos al momento de ser utilizados dentro de otro tipo de aplicaciones.

Además, al estudiar el test se descubren características peculiares en su rendimiento evaluado en términos de los errores de tipo I y II. Tales características se justifican por medio de la comprensión del fenómeno de separación del espectro, el cual gobierna el comportamiento del estadístico. En este sentido, el entendimiento del fenómeno pasa por el conocimiento y el análisis teórico del espectro de matrices aletorias, por lo que el estudio de la *RMT*, en este contexto, se vuelve indispensable.

Por otra parte, gracias al desarrollo e implementación del *Procedimiento de Factores* se logra darle una utilidad práctica al test *GOE*. Es más, el procedimiento, como producto final de la propuesta presentada en este trabajo, resulta ser un método válido de estimación del número de factores de una muestra. Esto se fundamenta en los resultados positivos de las pruebas empíricas a las que fue sometido en las sección 4 del capítulo 3.

En añadidura, el someter el procedimiento a experimentos, particularmente a aquellos que buscaban estudiar el error en la estimación, permitió fijar nociones acerca del comportamiento del estadístico. De esta forma, la naturaleza experimental de este trabajo fue predominante.

Otra conclusión importante es que el estudio del error en la estimación por parte del procedimiento, no solo sirvió para corroborar la teoría, sino también para establecer una pauta de las condiciones y parámetros bajo los cuales el procedimiento entrega resultados plausibles. Vale decir, considerar la regla entregada para las diferencias de los valores propios de la matriz de covarianza y el esquema de transición de fase para el α, encaminan la obtención de resultados a buen puerto.

Con respecto a las extensiones de este trabajo se vislumbra el estudio de la factibilidad de aplicación del procedimiento a datos reales. Es decir, dilucidar cuáles son los supuestos que

no se cumplen para el contexto de datos reales y evaluar la opción de levantarlos. En tal caso, se requeriría un análisis teórico para sustituirlos por otros más débiles o derechamente eliminarlos. Esto está fuertemente relacionado con la profundización del análisis de sensibilidad al que puede ser sometido el procedimiento cuando se perturban algunos de los supuestos del modelo.

Ahora bien, una línea de trabajo más inmediata sería continuar con la experimentación utilizando datos simulados con correlación. Específicamente, estudiar cómo cambia el comportamiento empírico del test GOE en términos de los errores que comete, y cómo esto podría repercutir en el procedimiento cuando existe correlación en los términos de idiosincracia. Otra alternativa que permite seguir esta dirección, es la implementación de experimentos que pongan a prueba al procedimiento con datos simulados de carácter dinámico. Es decir, podría desarrollarse el test, y por ende el procedimiento, bajo el modelo dinámico de factores definido en 2.1. Sin duda, estas últimas dos propuestas aportarían mayor realismo a las pruebas a las que está sujeto el test y el procedimiento.

Bibliografía

[1] Gernot Akemann, Jinho Baik, and Philippe Di Francesco. *The Oxford handbook of random matrix theory*. Oxford University Press, 2011.

[2] Ludwig Arnold. On the asymptotic distribution of the eigenvalues of random matrices. *Journal of Mathematical Analysis and Applications*, 20(2):262–268, 1967.

[3] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[4] Jinho Baik, Patrik L Ferrari, and Sandrine Péché. Limit process of stationary tasep near the characteristic line. *Communications on Pure and Applied Mathematics*, 63(8):1017–1070, 2010.

[5] Mireille Capitaine, Catherine Donati-Martin, Delphine Féral, et al. Central limit theorems for eigenvalues of deformations of wigner matrices. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 107–133. Institut Henri Poincaré, 2012.

[6] Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets, 1982.

[7] Ricardo Coelho, Peter Richmond, Stefan Hutzler, and Brian Lucey. A random-matrix-theory-based analysis of stocks of markets from different countries. *Advances in Complex Systems*, 11(05):655–668, 2008.

[8] Ivan Corwin, Patrik L Ferrari, and Sandrine Péché. Limit processes for tasep with shocks and rarefaction fans. *Journal of Statistical Physics*, 140(2):232–267, 2010.

[9] Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, 82(4):540–554, 2000.

[10] Josselin Garnier. Use of random matrix theory for target detection, localization, and reconstruction. *Contemporary Mathematics*, 548:1–19, 2011.

[11] Ulf Grenander. *Probabilities on algebraic structures*. Courier Corporation, 2008.

[12] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components

analysis. *Annals of statistics*, pages 295–327, 2001.

[13] Noureddine El Karoui. Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability*, pages 663–714, 2007.

[14] M Krbalek and P Seba. The statistical properties of the city transport in cuernavaca (mexico) and random matrix ensembles. *arXiv preprint nlin/0001015*, 2000.

[15] Alexei Onatski. A formal statistical test for the number of factors in the approximate factor models. *Department of Economics, Columbia University, Unpublished Manuscript*, 2006.

[16] Alexei Onatski. The Tracy-Widom limit for the largest eigenvalues of singular complex wishart matrices. *The Annals of Applied Probability*, pages 470–490, 2008.

[17] Alexei Onatski. Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479, 2009.

[18] James H Stock and Mark W Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002.

[19] Craig A Tracy and Harold Widom. Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, 159(1):151–174, 1994.

[20] Craig A Tracy and Harold Widom. The distributions of random matrix theory and their applications. In *New Trends in Mathematical Physics*, pages 753–765. Springer, 2009.

[21] Yong Q Yin. Limiting spectral distribution for a class of random matrices. *Journal of multivariate analysis*, 20(1):50–68, 1986.

Anexos

En este apartado se encuentran los anexos correspondientes al capítulo 3. Se trata de tablas de valores críticos e implementaciones en lenguaje Matlab de tests y procedimientos que, por economía visual, no fueron incluídos directamente.

Anexo A: Tabla de Valores Críticos *GUE*

A continuación se presenta la tabla de valores críticos para el estadístico complejo $\hat{R}$ de acuerdo al tamaño nominal del test *GUE*.

Tabla 3.3: Valores Críticos para el estadístico $\hat{R}$

	$k_1 - k_0$							
Tamaño test	1	2	3	4	5	6	7	8
15	2.78	3.65	4.16	4.57	4.90	5.15	5.42	5.65
10	3.37	4.34	4.93	5.36	5.76	6.03	6.35	6.60
9	3.54	4.54	5.13	5.59	5.99	6.28	6.60	6.89
8	3.71	4.76	5.37	5.86	6.24	6.56	6.92	7.19
7	3.91	5.01	5.66	6.13	6.57	6.92	7.28	7.57
6	4.22	5.32	6.01	6.48	6.98	7.35	7.71	8.01
5	4.56	5.71	6.41	6.98	7.49	7.84	8.21	8.56
4	4.97	6.19	7.01	7.56	8.09	8.51	8.94	9.24
3	5.56	6.90	7.81	8.47	9.03	9.40	9.81	10.14
2	6.49	8.05	9.04	9.66	10.18	10.67	11.25	11.70
1	8.35	10.02	11.21	12.12	13.01	13.75	14.63	15.09

Anexo B: Tabla de Valores Críticos *GOE*

A continuación se presenta la tabla de valores críticos para el estadístico R de acuerdo al tamaño nominal del test *GOE*.

Tabla 3.4: Valores Críticos para el estadístico R

	$k_1 - k_0$							
Tamaño test	1	2	3	4	5	6	7	8
15	3.58	5.23	6.24	7.07	7.88	8.63	9.28	9.89
10	4.65	6.60	7.84	8.96	9.98	10.90	11.67	12.36
9	4.93	6.97	8.33	9.52	10.56	11.58	12.32	13.10
8	5.26	7.43	8.88	10.12	11.33	12.26	13.16	13.94
7	5.71	8.01	9.54	10.92	12.12	13.17	14.10	15.06
6	6.21	8.74	10.38	11.82	13.20	14.30	15.33	16.33
5	6.93	9.71	11.44	13.06	14.56	15.79	16.94	17.99
4	7.84	10.89	12.82	14.70	16.43	17.73	18.88	20.17
3	9.26	12.65	14.94	17.13	18.99	20.40	22.03	23.53
2	11.55	15.69	18.20	21.27	23.56	25.56	27.64	29.39
1	16.32	22.38	26.35	31.24	33.66	36.63	39.46	42.33

Anexo C: Implementación en lenguaje Matlab del Test *GOE*

A continuación se muestra el código que implementa el test *GOE* como una función en lenguaje Matlab.

```
%%%%TEST H0:k=k0 CONTRA H1: k0 <k <=k1 FACTORES

%%%%Recibe data de la forma TiempoxVariables
%%%%Entrega un vector columna con el resultado del test para cada tamaño.
%%%%Luego, para un tamaño en particular se debe seleccionar
%%%%el elemento correspondiente del vector.

function out=TestGOE(Data, k0, k1)

%%% Inputs: Data—Data a la que se le quiere aplicar el test
%%%         k0—número de factores que se quiere testear
%%%         k1—número máximo de factores
%%% Ouput: Resultado del test de k=k0 factores versus k0 <k <=k1 factores.
%%%        Entrega un 1 si se rechaza la hipótesis nula y un 0 si no.

Data=Data';                    %Data de dimensiones nxT
[n, T]=size(Data);
MatrizCov=(1/T)*(Data*Data');  %Matriz de covarianza
opts.disp=0;                   %Opciones del display
vp=eigs(MatrizCov, k1+2,'la', opts); %Valores propios matriz de covarianza

%Valor del estadístico R
R=max((vp(k0+1:k1)—vp(k0+2:k1+1))./(vp(k0+2:k1+1)—vp(k0+3:k1+2)));

%Regla de decisión
load ValoresCriticosGOE.txt      %Carga tabla de valores críticos
[m h]=size(ValoresCriticosGOE); %Dimensiones tabla de valores críticos

RESULTADOGUE_Ec=zeros(m,1); %Inicia variable Resultados
for k=1:m                   %Ciclo for resultado del test para cada alpha
    if R> ValoresCriticosGOE(k,k1—k0)
        RESULTADOGUE_Ec(k)=1; %Se codifica con 1 si rechaza H0
    else
        RESULTADOGUE_Ec(k)=0; %Se codifica con 0 si no rechaza H0
    end
end

out=RESULTADOGUE_Ec;          % Output: Resultado binario para cada alpha

end
```

Anexo D: Implementación en lenguaje Matlab del Procedimiento de Factores

A continuación se muestra el código que implementa el *Procedimiento de Factores* como una función en lenguaje Matlab.

```
%%%%%PROCEDIMIENTO PARA DETERMINAR EL NÚMERO DE FACTORES
%%%%%Retorna el primer k0 que el test no rechaza

%%%% Inputs: Data: Data a la que se le quiere aplicar el test
%%%%         k1: nro máximo de factores

%%%% Output: El primer k0 que no es rechazado

function out=ProcFactores(Data, k1)

out=[];              %Inicializa variable output

   for k0=1:(k1-1)   %Ciclo for para iterar los tests
        t=TestGOE(Data,k0,k1);   %Aplica el test para k0
        if t(7)==0;              %Se considera el tamaño nonimal del 5%
        out=k0;                  %Si no rechazó k0, este es el output
        break
        end

        if(isempty(out)==1)  %Si rechaza todos los valores del ciclo
        out=k1;              %entonces, el output es k1
        end
   end


end
```

Anexo E: Tasas de pérdida de coherencia del *Procedimiento de Factores*

A continuación se presenta la tabla de tasas de pérdida de coherencia por parte del *Procedimiento de Factores*. La tabla es parte de los resultados discutidos al final de la sección 3.5 .

Las tasas se muestran para N muestras simuladas con $4, 5, \ldots, 12$ factores, sobre las cuales se aplicó el *Procedimiento de Factores* con un parámetro $\alpha = 5\,\%$. Además la tasa de pérdida promedio que se obtiene es de 0,4836 %.

Tabla 3.5: Tasas Pérdida de Coherencia

N	Factores en la muestra	Tasa (%)
10000	4	0.3175
10000	5	0.3694
10000	6	0.3058
10000	7	0.5374
10000	8	0.5071
10000	9	0.5048
10000	10	0.5050
10000	11	0.6310
10000	12	0.6743