UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ

INSTITUTO DE FÍSICA

DESARROLLO DE ESTRATEGIAS DE PAIR TRADING MEDIANTE COINTEGRACIÓN BASADA EN MATRICES ALEATORIAS

T E S I S

PARA OBTENER EL GRADO DE:

MAESTRÍA EN CIENCIAS (FÍSICA)

PRESENTA:

LIC. SERVANDO FARID ABDALA MARTÍNEZ

DIRECTOR DE TESIS:

DR. JOHN ALEXANDER FRANCO VILLAFAÑE

CO-DIRECTOR DE TESIS:

DR. ANDRÉS GARCÍA MEDINA

SAN LUIS POTOSÍ, S.L.P., MÉXICO

18 de Septiembre del 2026

Desarrollo de Estrategias de Pair Trading Mediante Cointegración Basada en Matrices Aleatorias

A mi familia, máquina de movimiento perpetuo en esta caminata aleatoria llamada vida.

A mi madre, Martha Beatriz Martínez Flores: gracias por tu apoyo incondicional y por mostrarme el verdadero significado del amor.

A mi padre, Servando Abdala Saldaña: tu ejemplo y respaldo como padre son el modelo de todo lo que anhelo ser para mis hijos.

A mi hermana, Jade Sarahí Abdala Martínez: eres mi mayor motivación para superarme y ser tu referente. Te amo profundamente.

Este logro es también de ustedes; es tan suyo como mío.

AGRADECIMIENTOS

Quisiera expresar mi más profundo agradecimiento a mis directores de tesis, el Dr. John Alexander Franco Villañañe y el Dr. Andrés García Medina, por su invaluable guía, su infinita paciencia y la generosidad con que compartieron su sabiduría. Más que tutores académicos, fueron faros en los momentos de incertidumbre, ayudándome a transformar cada obstáculo en oportunidad de aprendizaje. Su confianza en mí y en este proyecto fue el mayor estímulo para seguir adelante.

A los profesores del Instituto de Física, por su dedicación y por sembrar en mí la curiosidad científica que hoy da frutos. Cada clase, cada conversación, cada consejo quedará grabado como parte fundamental de mi formación.

A mis compañeros y amigos de carrera y maestría, por las incontables horas de estudio compartidas y el apoyo mutuo en este viaje.

A mi familia, pilar fundamental de mi existencia. A mis padres, por su amor sin condiciones, por cada sacrificio que hicieron para que yo pudiera llegar hasta aquí, por enseñarme que los sueños se construyen con esfuerzo y honestidad. A mi hermana, por ser mi refugio y mi alegría, por recordarme siempre lo que realmente importa.

A todos los que, de una u otra forma, contribuyeron a que esta tesis sea realidad: gracias por ser parte de mi historia. Este logro no es solo mío, es de todos los que caminaron a mi lado, en cuerpo y mente.

¡GRACIAS... TOTALES!

Resúmen

La identificación de relaciones de cointegración en mercados financieros es equivalente a detectar modos colectivos de largo plazo en un sistema de muchos cuerpos. Desde la óptica de la física estadística, un mercado compuesto por N activos financieros constituye un sistema interactuante con N grados de libertad, cuya dinámica temporal está registrada en T observaciones. Los métodos clásicos de cointegración, como el test de Johansen, presentan limitaciones inherentes cuando N es grande: la matriz de covarianzas muestral se vuelve mal condicionada, sus autovalores se contaminan con ruido de fondo—descrito por la distribución de Marčenko-Pastur—y la señal cointegrante se diluye en el bulbo espectral. Esta maldición de la dimensionalidad restringe el análisis a subportafolios de reducido tamaño, lo que implica un alto costo computacional, una pérdida de oportunidades de arbitraje en tiempo real y una escasa adaptabilidad a la llegada de nueva información.

Para superar estas limitaciones, Bikhovskaya et al. **bikhovskaya2021** propusieron el test *LargeVAR*, una reformulación del test de Johansen que incorpora herramientas de la Teoría de Matrices Aleatorias (RMT) y que resulta consistente incluso en regímenes de alta dimensionalidad ($N \sim T$ o $N > T$). Sin embargo, la literatura se ha concentrado en la *detección* del número de relaciones de cointegración, dejando de lado la *estimación* de las relaciones mismas, que es el paso necesario para cualquier aplicación práctica, como el *pair trading*.

La presente investigación aborda este vacío mediante una aproximación experimental y empírica. En primer lugar, se generan datos sintéticos a partir de un Modelo de Corrección de Error Vectorial (VECM) con relaciones de cointegración conocidas, lo que permite un entorno controlado de validación. Sobre estos datos, se aplican tanto el test de Johansen (baja dimensión) como el test *LargeVAR* (alta dimensión), evaluando su precisión en la detección del rango de cointegración. Como contribución central, se extiende la metodología *LargeVAR* para estimar no solo el número de relaciones, sino también los vectores de cointegración β y la matriz de ajuste α, parámetros fundamentales para la construcción de *spreads* operativos. Posteriormente, se exploran estrategias de *pair trading* basadas en estas estimaciones, evaluando su desempeño mediante métricas propias de la industria financiera.Es importante señalar que el objetivo fundamental de este trabajo es demostrar la viabilidad de identificar correctamente las relaciones de cointegración en alta dimensión; la rentabilidad es una métrica secundaria que permite validar la utilidad práctica de dicha identificación. Finalmente se contrastan los resultados con datos empíricos de mercados reales para verificar la aplicabilidad del enfoque. Toda la implementación se realiza en Python, traduciendo y extendiendo el código original en R, con el doble

objetivo de validar los resultados del trabajo seminal y de democratizar el acceso a estas herramientas mediante una librería de código abierto.

De este modo, la tesis no solo ofrece una validación independiente del test *LargeVAR*, sino que extiende su alcance para incluir la estimación de los parámetros clave del modelo de cointegración, proponiendo un puente metodológico entre la detección en alta dimensión y su aplicación práctica, enmarcado en el lenguaje y los principios de la econofísica.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

1.1. Contexto general

La econofísica es una de esas disciplinas híbridas que más me atraen, por no decir que es la que mas me llama la atención especialmente porque toma herramientas de la física estadística y las aplica a fenómenos económicos y financieros [1]. En este campo, cuando hablamos de series temporales financieras nos referimos a secuencias de observaciones de una variable (pongamos el precio de un activo) espaciadas uniformemente en el tiempo. Si tengo N activos y T períodos, puedo organizar los datos en una matriz de $N \times T$: cada fila recoge la historia de un activo, y cada columna, la instantánea de todos los activos en un momento dado.

El análisis multivariante de estas series trata de sacar a la luz las relaciones ocultas entre los activos. Y aquí la cointegración merece una mención especial: básicamente, nos dice que hay un equilibrio de largo plazo entre series que, por sí solas, no son estacionarias [2]. En términos formales, dos o más series integradas de orden 1 ($I(1)$) están cointegradas si existe alguna combinación lineal de ellas que sí es estacionaria ($I(0)$). Esa combinación se mueve alrededor de una media constante; las desviaciones son, por tanto, temporales y esto es lo interesante para un inversor ofrecen oportunidades basadas en la reversión a la media. Las estrategias de *pair trading* [3] son el ejemplo clásico.

1.2. Problema de investigación

Para sacar a la luz estas relaciones de cointegración que originalmente se encuentran ocultas, la prueba de Johansen [4], [5] se ha convertido en el estándar a la hora de contar cuántas relaciones de cointegración existen en un sistema multivariante. Asintóticamente, con N fijo y T tendiendo a infinito, el método es consistente y

óptimo. Pero la realidad es tozuda: los fondos de inversión manejan carteras con cientos de activos, un escenario de alta dimensión donde N y T son perfectamente comparables. En esas condiciones, los métodos clásicos empiezan a fallar estrepitosamente: las matrices de covarianza muestral se vuelven mal condicionadas y tienden a encontrar relaciones espurias [6]. En la práctica, uno se ve obligado a partir el universo de activos en subconjuntos más pequeños y a realizar múltiples pruebas, lo que genera un alto costo computacional.

Bykhovskaya y Gorin [7] en su trabajo lograron modificar la prueba de Johansen usando teoría de matrices aleatorias (RMT), tranformado este problema en una oportunidad. El resultado, bautizado como *LargeVARS*, consigue determinar el número de relaciones de cointegración incluso cuando N y T son del mismo orden, esquivando el sobre rechazo que arruina las pruebas tradicionales en alta dimensión.

Sin embargo (y aquí empieza el verdadero reto), saber el rango r no le basta al operador financiero. Con N activos, las combinaciones lineales posibles son un océano, y lo que de verdad importa es señalar con el dedo qué pares (o grupos) concretos de activos están cointegrados y cuyas desviaciones del equilibrio son explotables. Así que el problema se desplaza: hay que construir una metodología que, apoyándose en la detección robusta del rango que ofrece *LargeVARS*, sea capaz de (i) identificar los pares cointegrados, (ii) estimar sus vectores de cointegración y (iii) utilizar esa información para montar estrategias de *pair trading*.

1.3. Justificación

A mi modo de ver, hay una brecha importante entre los avances teóricos en la detección de cointegración en alta dimensión y su uso práctico. El trabajo de Bykhovskaya y Gorin [7] resuelve el problema de la detección con ajustes basados en RMT, pero no va más allá: no nos dice cómo aprovechar esa información para diseñar estrategias de inversión. Mientras tanto, el *pair trading* tradicional se ha apoyado en la selección de pares por correlación o en pruebas de cointegración en baja dimensión. El problema es que la correlación no garantiza una relación de equilibrio a largo plazo, y las pruebas de cointegración clásicas, en entornos ruidosos, son frágiles [8].

Esta tesis intenta tender un puente entre la teoría y la práctica. Lo que propongo es extender *LargeVARS* hasta convertirla en una metodología completa: desde la detección robusta de cointegración hasta la construcción y evaluación de estrategias de *pair trading*. Que los resultados finales sean espectaculares o modestos, creo que el ejercicio tiene valor académico: nos ayuda a entender mejor cómo funciona la cointegración en alta dimensión y cómo se puede canalizar hacia decisiones de

inversión.

1.4. Pregunta de investigación y objetivos

La pregunta que me hago, en última instancia, es si se puede diseñar una metodología que, con pruebas de cointegración basadas en matrices aleatorias, genere estrategias de *pair trading* con el riesgo bajo control. De ahí surgen varias preguntas más concretas:

- ¿Cómo adaptar *LargeVARS* para que, más allá de decirme cuántas relaciones de cointegración hay, me permita señalar pares cointegrados y estimar sus vectores?
- Frente al Johansen de toda la vida, ¿gana este enfoque en potencia de detección y en precisión cuando N y T son comparables?
- Y, en la práctica: ¿qué rentabilidad y qué riesgo presentan las estrategias construidas así, tanto sobre datos simulados como sobre el mercado real de acciones estadounidense?

El objetivo general, por tanto, es desarrollar y poner a prueba una metodología basada en *LargeVARS* que sea capaz de generar estrategias de *pair trading* en alta dimensión, identificando pares cointegrados y estimando sus ecuaciones de cointegración.

En concreto, me propongo:

1. Desarrollar una librería en Python que implemente *LargeVARS* y el método de Johansen, partiendo del código original en R y añadiendo funcionalidades que faciliten su uso a otros investigadores.
2. Construir un entorno de simulación Monte Carlo basado en un modelo VECM con propiedades de cointegración controladas, que permita comparar con justicia la potencia para detectar r y la precisión del vector de cointegración de ambos métodos.
3. Dentro de ese entorno, comparar a fondo *LargeVARS* y Johansen, con especial atención a los escenarios de alta dimensión (N grande y T/N moderado).
4. Diseñar un procedimiento que, a partir de la salida de *LargeVARS*, seleccione pares cointegrados y estime sus vectores, corrigiendo los sesgos de muestras finitas cuando sea necesario.

5. Medir la rentabilidad y el riesgo de las estrategias de *pair trading* resultantes mediante un *backtesting* con datos históricos del S&P500, apoyado en simulaciones para asegurar la robustez de las conclusiones.

1.5. Estructura de la tesis

He organizado la tesis en siete capítulos. Después de esta introducción, en el Capítulo 2 repaso los fundamentos del análisis de series temporales, la cointegración, el modelo VECM y la prueba de Johansen, para sentar las bases. El Capítulo 3 se adentra en la teoría de matrices aleatorias (RMT) aplicada a la cointegración: el ensamble de Jacobi, las distribuciones de Wachter y Tracy-Widom, y la prueba *LargeVARS*. Con esos mimbres, el Capítulo 4 presenta un estudio de simulación donde pongo a prueba la potencia y precisión de *LargeVARS* frente a Johansen en escenarios de alta dimensión. El Capítulo 5 detalla cómo extiendo *LargeVARS* para identificar pares cointegrados y estimar sus vectores. El Capítulo 6 explica los fundamentos del *pair trading* y cómo, a partir de los resultados anteriores, construyo estrategias concretas con reglas de entrada, salida y gestión del riesgo. Finalmente, el Capítulo 7 recoge los resultados empíricos, responde a las preguntas de investigación y cierra con conclusiones y posibles líneas futuras.

Además, el apéndice incluye la dirección del repositorio público donde está la librería de Python desarrollada, con todo el código de *LargeVARS* y los scripts de simulaciones y *backtesting*, para que cualquiera pueda reproducir los resultados.

Capítulo 2

Análisis de Series de Tiempo Financieras para Detectar Cointegración

2.1. Análisis de Series Temporales

El análisis de series temporales es una de esas disciplinas que, cuando la estudias a fondo, te das cuenta de que está en todas partes. Originalmente arraigada en la estadística y la econometría, su utilidad desborda con creces el ámbito financiero: aparece en geociencias, climatología, ingeniería y, por supuesto, en el estudio de sistemas complejos dentro de la física [9]-[11]. En finanzas, desde luego, es el pan de cada día: modelar precios, volatilidad o relaciones de cointegración no se entiende sin ella [12], [13].

2.1.1. Definición y Características

Llamamos serie temporal a cualquier secuencia de observaciones $\{y_t\}_{t=1}^{T}$ ordenadas en el tiempo. Si uno quiere ser más preciso, puede pensarla como la realización concreta de un proceso estocástico subyacente $\{Y_t : t \in \mathcal{T}\}$, donde $\mathcal{T}$ es un conjunto de índices temporales [10], [13]. Esta distinción entre la serie observada y el proceso que la genera es, a mi juicio, crucial para no malinterpretar los resultados.

Cuando descomponemos una serie en sus piezas, el esquema clásico sigue siendo una referencia útil:

$$Y_t = T_t + S_t + C_t + I_t, \tag{2.1}$$

donde T_t es la tendencia (ese movimiento suave de fondo), S_t recoge los patrones

estacionales que se repiten con periodicidad fija, C_t refleja los ciclos —más erráticos, de medio plazo— y I_t es el residuo, el ruido no sistemático que no podemos atribuir a nada concreto. En la Figura 2.1 se ve este ejercicio aplicado al precio del petróleo mexicano; saltan a la vista las distintas capas que componen la serie original.

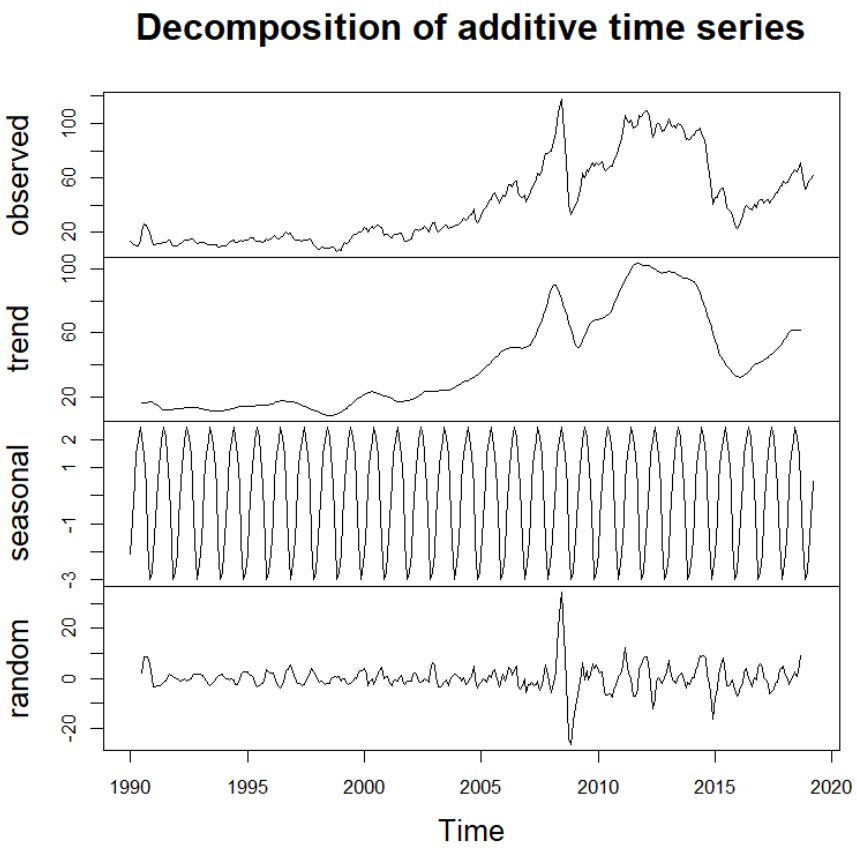


Figura 2.1: Descomposición de la serie de precios del petróleo mexicano en sus componentes: tendencia, estacionalidad y residuales.

2.1.2. Notación Matricial

Cuando pasamos al análisis multivariado —y en particular al modelo VECM de la Sección 2.8—, conviene soltar la notación escalar y adoptar el lenguaje del álgebra lineal. No es capricho: las matrices nos dan claridad visual y herramientas operativas potentes [14], [15].

Si tenemos N activos y T observaciones, los datos caben en una matriz $\mathbf{X}$ de $N \times T$:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,T} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,T} \end{bmatrix}, \tag{2.2}$$

donde x_{it} es la observación de la serie i en el instante t. En finanzas, las filas suelen ser los activos (acciones, divisas...) y las columnas los días o semanas de la muestra.

Las transformaciones más habituales con esta matriz son tres:

1. **Matriz de retornos:** A partir de los precios $P_{i,t}$, los retornos logarítmicos son

$$R_{i,t} = \ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right), \quad \mathbf{R} \in \mathbb{R}^{N\times(T-1)}. \tag{2.3}$$

2. **Matriz de covarianza:** Asumiendo retornos de media cero, la matriz de covarianza muestral se calcula como

$$\mathbf{S} = \frac{1}{T-1}\mathbf{R}\mathbf{R}^{\top}. \tag{2.4}$$

3. **Matriz de correlación:** Basta con estandarizar:

$$\mathbf{C} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}, \tag{2.5}$$

donde $\mathbf{D} = \mathrm{diag}(s_{11}, \ldots, s_{NN})$ contiene las varianzas individuales.

El verdadero quebradero de cabeza aparece cuando N es grande. Si la matriz de covarianza muestral se vuelve singular o está mal condicionada, las estimaciones se llenan de ruido y cualquier inferencia posterior se resiente [6]. Volveremos a esto más adelante.

2.2. Estacionariedad

Si tuviera que elegir un concepto que es la columna vertebral de todo el análisis de series temporales, ese sería la estacionariedad. Las condiciones estrictas rara vez se cumplen en la práctica, así que manejamos dos definiciones complementarias [10], [13].

La **estacionariedad estricta** (o fuerte) exige que la distribución conjunta de cualquier subconjunto de observaciones sea invariante a desplazamientos en el tiempo. Formalmente, para cualquier conjunto de índices $t_1, t_2, \ldots, t_k$ y cualquier h:

$$F_{Y_{t_1},Y_{t_2},\ldots,Y_{t_k}}(y_1, y_2, \ldots, y_k) = F_{Y_{t_1+h},Y_{t_2+h},\ldots,Y_{t_k+h}}(y_1, y_2, \ldots, y_k). \tag{2.6}$$

Pero como digo, en la práctica eso es pedir demasiado. La **estacionariedad débil** (o de segundo orden) se conforma con tres condiciones mucho más manejables: media constante, varianza finita y constante, y autocovarianza que dependa solo del rezago h, no del tiempo t. Si un proceso $\{Y_t\}$ satisface

$$\mathbb{E}[Y_t] = \mu, \quad \text{Var}(Y_t) = \sigma^2 < \infty, \quad \text{Cov}(Y_t, Y_{t+h}) = \gamma(h),$$

para todo t y todo entero h, entonces decimos que es débilmente estacionario. La Figura 2.2 muestra un ejemplo donde se aprecian esas constantes.

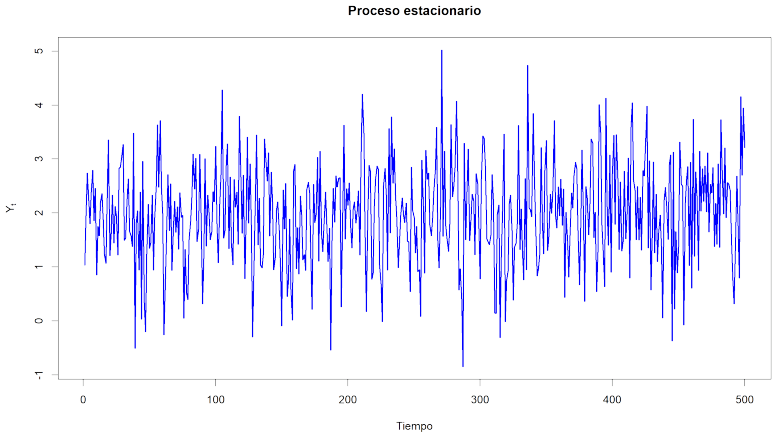


Figura 2.2: Ejemplo de un proceso estacionario: media y varianza constantes, y estructura de autocorrelación que depende solo del rezago.

2.3. Problema de Raíz Unitaria

Buena parte de las series económicas y financieras tienen un problema: no son estacionarias, y con frecuencia la causa es una raíz unitaria. Dicho rápido, una serie $\{y_t\}$ tiene una raíz unitaria si el polinomio característico de su representación autorregresiva tiene una raíz exactamente igual a uno [10], [13].

Pensemos en un AR(p):

$$\phi(B)y_t = \varepsilon_t, \tag{2.7}$$

con $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$. Si $\phi(1) = 0$, hay raíz unitaria. Y eso tiene consecuencias prácticas de mucho calado: la serie deja de ser estacionaria —su

media y varianza dependen de t—, su varianza crece sin cota, los *shocks* tienen efectos permanentes (memoria infinita) y, por si fuera poco, las regresiones entre series con raíces unitarias pueden resultar espurias: coeficientes que parecen significativos pero que solo reflejan tendencias comunes, no relaciones genuinas [16].

2.3.1. Orden de Integración

Decimos que y_t es integrada de orden d, $y_t \sim I(d)$, si hay que diferenciarla d veces para volverla estacionaria. Así, $I(0)$ es estacionaria en niveles, $I(1)$ necesita una diferencia, $I(2)$ dos, y así sucesivamente.

2.3.2. Diferenciación

La herramienta más directa para eliminar una raíz unitaria es la diferenciación. El operador de primera diferencia es simple:

$$\Delta y_t = y_t - y_{t-1} = (1 - B)y_t. \tag{2.8}$$

Si con eso basta para obtener una serie estacionaria, estamos ante una $I(1)$.

2.4. Modelos Univariados Fundamentales

Antes de meternos con varios activos a la vez, conviene tener claros los ladrillos básicos del análisis univariado [10], [13]. Estos modelos son el punto de partida.

2.4.1. Proceso de Ruido Blanco

Una secuencia $\{\varepsilon_t\}$ es ruido blanco si cumple tres condiciones: media cero, varianza finita y constante, y ausencia total de correlación serial. Es el bloque más simple —y, sin embargo, el más ubicuo— porque aparece como innovación en prácticamente todos los modelos que veremos. La Figura 2.3 muestra un ejemplo generado a partir de una normal estándar.

2.4.2. Modelo Autorregresivo (AR)

Un AR(p) tiene esta pinta:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \tag{2.9}$$

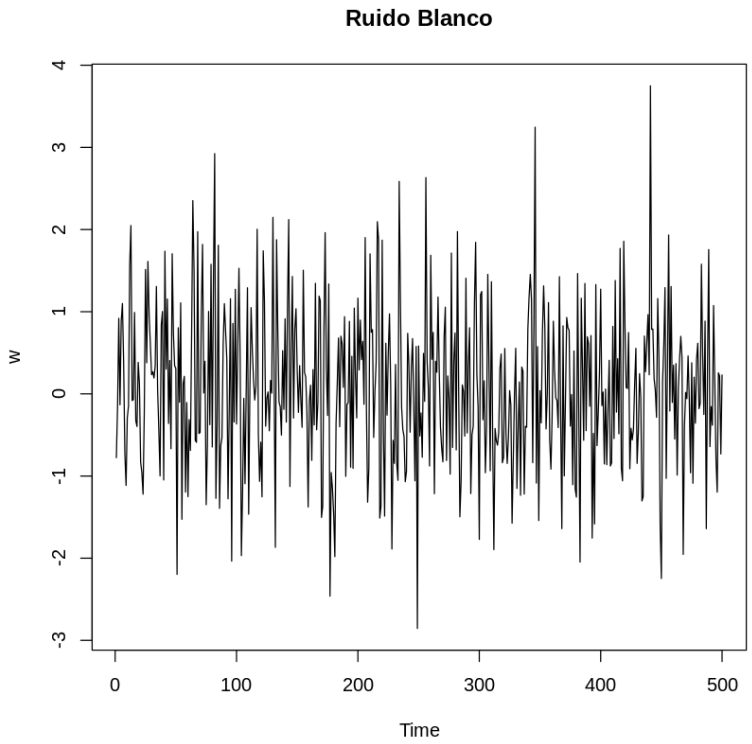


Figura 2.3: Serie de ruido blanco generada a partir de una distribución normal estándar.

con $\varepsilon_t \sim \text{WN}(0, \sigma^2)$. La variable se explica por su propio pasado. La Figura 2.4 muestra un AR(2) simulado: se ven bien las oscilaciones que lo caracterizan.

2.4.3. Modelo de Media Móvil (MA)

Un MA(q) explica y_t mediante innovaciones pasadas:

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}. \tag{2.10}$$

Aquí μ es la media del proceso. La Figura 2.5 ilustra su aspecto típico.

2.4.4. Modelos ARMA y ARIMA

Combinar AR y MA da lugar al ARMA(p, q):

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}. \tag{2.11}$$

Y si la serie no es estacionaria pero sus diferencias sí, hablamos de ARIMA(p, d, q):

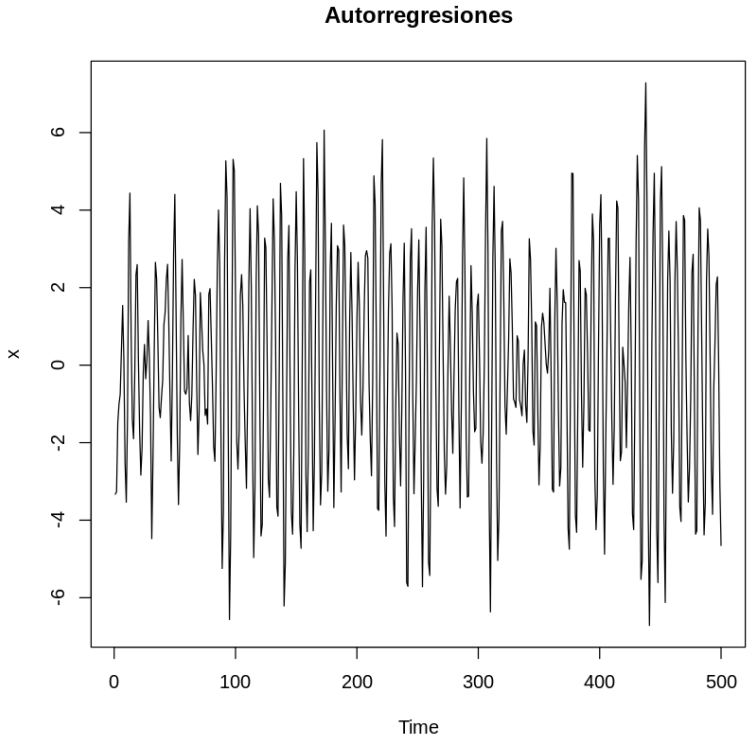


Figura 2.4: Ejemplo de un proceso AR(2) simulado.

$$\phi(B)(1-B)^d y_t = c + \theta(B)\varepsilon_t. \tag{2.12}$$

La d indica cuántas veces hemos tenido que diferenciar para alcanzar la estacionariedad.

2.5. Funciones de Dependencia Temporal

Cuando empecé a trabajar con series financieras, pronto me di cuenta de que sin estas funciones uno avanza a ciegas. Son, sencillamente, la columna vertebral del diagnóstico [10], [13]. Si no entiendes cómo se relaciona una serie consigo misma o con otras, difícilmente podrás modelarla bien.

2.5.1. Autocovarianza

En un proceso débilmente estacionario $\{y_t\}$ con media μ, la autocovarianza de orden h es la covarianza entre la serie y ella misma desplazada h periodos:

$$\gamma(h) = \mathrm{Cov}(y_t, y_{t-h}) = \mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)]. \tag{2.13}$$

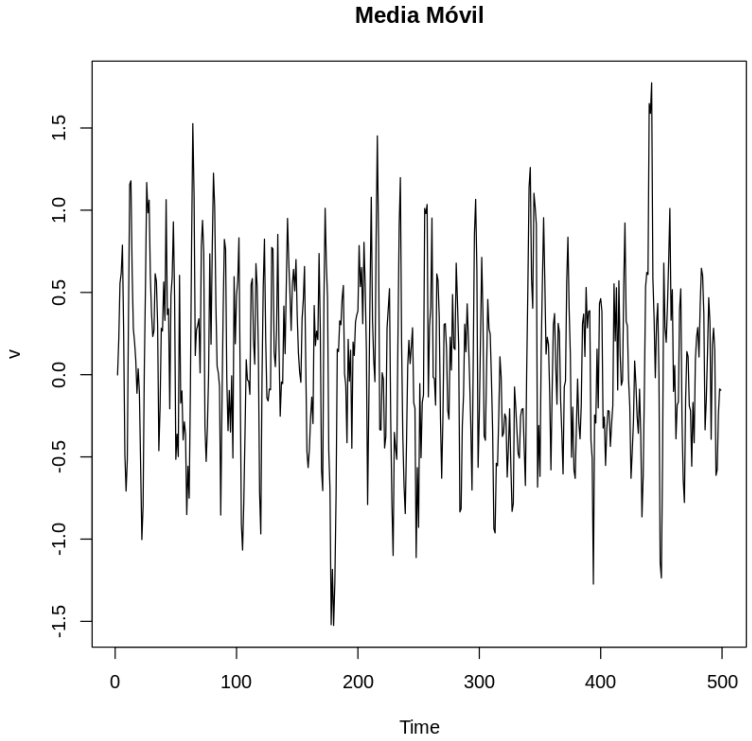


Figura 2.5: Ejemplo de un proceso de media móvil.

2.5.2. Autocorrelación (ACF)

La ACF no es más que la autocovarianza normalizada para que tome valores entre -1 y 1:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \text{Corr}(y_t, y_{t-h}). \tag{2.14}$$

2.5.3. Autocorrelación Parcial (PACF)

La PACF va un paso más allá: mide la correlación entre y_t e y_{t-h} una vez que hemos descontado el efecto de todos los rezagos intermedios. Formalmente, $\phi_{h,h}$ es el coeficiente de y_{t-h} en la regresión de y_t sobre sus h rezagos:

$$y_t = \phi_{h1} y_{t-1} + \phi_{h2} y_{t-2} + \cdots + \phi_{hh} y_{t-h} + \varepsilon_t^{(h)}. \tag{2.15}$$

2.5.4. Correlación Cruzada (CCF)

Cuando en lugar de una serie tenemos dos procesos estacionarios $\{x_t\}$ e $\{y_t\}$, la CCF captura la relación entre x_t y y_{t-h}:

$$\rho_{xy}(h) = \text{Corr}(x_t, y_{t-h}) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}, \tag{2.16}$$

con $\gamma_{xy}(h) = \text{Cov}(x_t, y_{t-h})$.

2.6. Modelos Multivariados

Cuando tenemos un vector de k series que se influyen entre sí, los modelos univariados se quedan cortos. Necesitamos modelos multivariados. Son la antesala natural de la cointegración [5], [12].

2.6.1. Modelo de Vectores Autorregresivos (VAR)

Un VAR(p) para $\mathbf{y}_t = (y_{t1}, \ldots, y_{tk})^\top$ generaliza el AR al caso vectorial: cada variable se explica por su propio pasado y por el pasado de todas las demás.

$$\mathbf{y}_t = \boldsymbol{\nu} + \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t, \tag{2.17}$$

donde $\boldsymbol{\nu}$ es un vector de constantes, $\mathbf{A}_i$ son matrices $k \times k$ de coeficientes y $\mathbf{u}_t$ el vector de innovaciones.

2.7. Cointegración

El trabajo de Engle y Granger [2] es donde se define el concepto de cointegración, en retrospectiva, resultan casi obvia para analizar series temporales. Por fin había un marco riguroso para modelar relaciones de equilibrio de largo plazo entre series no estacionarias.

2.7.1. Definición Formal y Condiciones

Partimos de un vector $\mathbf{y}_t$ con k series, cada una $I(d)$. Decimos que están cointegradas de orden (d, b) si se cumplen dos condiciones:

1. Todas comparten el mismo orden de integración d.
2. Existe al menos un vector $\boldsymbol{\beta} \neq \mathbf{0}$ tal que la combinación lineal

$$z_t = \boldsymbol{\beta}^\top \mathbf{y}_t = \beta_1 y_{1t} + \beta_2 y_{2t} + \cdots + \beta_k y_{kt} \tag{2.18}$$

resulta ser $I(d-b)$ con $b > 0$.

En la práctica, el caso que nos ocupa es $d = 1, b = 1$: series $I(1)$ cuya combinación lineal es $I(0)$, es decir, estacionaria.

2.7.2. Prueba de Johansen

A mi juicio, la prueba de Johansen sigue siendo la herramienta más general y robusta para detectar y estimar relaciones de cointegración en sistemas de bajas dimensiones [4], [5], [17]. A diferencia de enfoques más limitados, permite identificar múltiples vectores de cointegración simultáneamente.

El núcleo de la prueba está en el rango de la matriz $\mathbf{\Pi}$ de la representación VECM. Si rango($\mathbf{\Pi}$) $= 0$, no hay cointegración. Si $0 < r < k$, existen r relaciones. Y si $r = k$, todas las series ya eran $I(0)$ desde el principio, así que mal planteamiento.

¿Cómo decidimos ese r? Johansen propone dos estadísticos basados en los autovalores $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_k$:

1. **Prueba de la traza**: La hipótesis nula es que hay como máximo r vectores de cointegración.
$$\lambda_{\text{traza}}(r) = -T \sum_{i=r+1}^{k} \ln(1 - \hat{\lambda}_i). \tag{2.19}$$

2. **Prueba del máximo autovalor**: Aquí contrastamos r frente a $r + 1$.
$$\lambda_{\max}(r, r+1) = -T \ln(1 - \hat{\lambda}_{r+1}). \tag{2.20}$$

¿Por qué dos estadísticos? Porque responden a preguntas ligeramente distintas: la traza evalúa el número total de relaciones, mientras que el máximo autovalor examina si añadir una relación más mejora el modelo de forma significativa. El procedimiento es secuencial: se empieza con $r = 0$; si se rechaza, se pasa a $r = 1$, y así hasta que no se pueda rechazar la nula. Una vez fijado r, los vectores de cointegración $\boldsymbol{\beta}$ y las velocidades de ajuste $\boldsymbol{\alpha}$ se estiman por máxima verosimilitud.

Entre sus ventajas, destacaría que maneja múltiples relaciones a la vez, utiliza toda la información del sistema, permite contrastar restricciones y es eficiente bajo normalidad. Nada despreciable.

2.8. Modelo de Corrección de Error Vectorial (VECM)

El VECM no es más que la forma natural de escribir un VAR cuando existe cointegración [5]. Separa la dinámica de corto plazo del ajuste hacia el equilibrio de largo plazo. Si partimos de un VAR(p) para $\mathbf{y}_t$ ($I(1)$), lo reescribimos como:

$$\Delta \mathbf{y}_t = \boldsymbol{\nu} + \boldsymbol{\Pi} \mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \Delta \mathbf{y}_{t-i} + \mathbf{u}_t. \tag{2.21}$$

Incluyendo términos determinísticos, la forma más general es:

$$\Delta \mathbf{y}_t = \boldsymbol{\alpha} \boldsymbol{\beta}^\top \mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\Phi} \mathbf{D}_t + \boldsymbol{\varepsilon}_t. \tag{2.22}$$

2.8.1. La Matriz de Impacto

Bajo cointegración, $\boldsymbol{\Pi}$ tiene rango reducido r y se descompone en dos piezas con interpretación muy intuitiva:

$$\boldsymbol{\Pi} = \boldsymbol{\alpha} \boldsymbol{\beta}^\top, \tag{2.23}$$

donde $\boldsymbol{\beta}$ $(k \times r)$ contiene los vectores de cointegración y $\boldsymbol{\alpha}$ $(k \times r)$ las velocidades de ajuste. Dicho sin fórmulas: $\boldsymbol{\beta}^\top \mathbf{y}_{t-1}$ mide cuánto nos hemos desviado del equilibrio, y $\boldsymbol{\alpha}$ nos dice con qué intensidad reacciona cada variable para corregir esa desviación.

2.9. Correlación y Cointegración

Conviene no confundir estos dos conceptos: ambos hablan de relación entre series, pero la naturaleza de esa relación es completamente distinta [2]. La correlación mide asociación lineal sincrónica (o con desfase) y es muy sensible al período muestral. La cointegración, en cambio, captura una relación estructural de largo plazo.

Con series estacionarias, la ACF $\rho(h)$ de la ecuación (2.14) depende solo de h. Pero cuando las series son $I(1)$, la correlación muestral puede bailar de una ventana a otra sin que eso refleje nada profundo. ¿Significa eso que no hay relación? No necesariamente. La cointegración sí es una propiedad que perdura: dos series $I(1)$ están cointegradas si existe $\boldsymbol{\beta}$ tal que $\boldsymbol{\beta}^\top \mathbf{y}_t$ es $I(0)$. Esa combinación estacionaria actúa como un imán: las series pueden separarse en el corto plazo, pero hay una fuerza que las reconduce.

La Figura 2.6 lo muestra con claridad en cuatro paneles. Arriba a la izquierda, diez series $I(1)$ simuladas; dos de ellas (naranja y azul) están cointegradas, aunque a simple vista no se note. Arriba a la derecha, esas dos series aisladas: en subperíodos la correlación puede ser positiva, negativa o casi nula. Abajo a la izquierda, la combinación cointegrada $z_t = y_{1t} - \beta y_{2t}$ es claramente estacionaria. Abajo a la derecha, las diferencias de las series confirman que ambas son $I(1)$.

En definitiva, que dos activos no muestren una correlación estable no significa

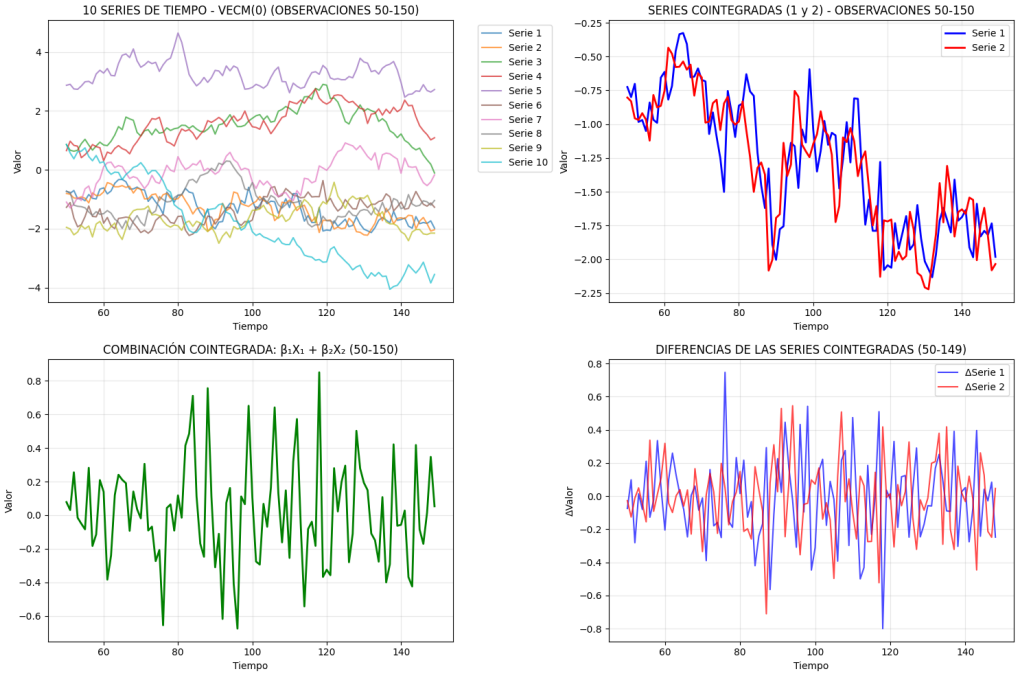


Figura 2.6: Ilustración de la relación entre correlación y cointegración. (Arriba-izquierda): Diez series I(1) simuladas, con dos de ellas (naranja y azul) cointegradas. (Arriba-derecha): Las dos series cointegradas, mostrando correlación variable en sub-períodos pero movimiento conjunto a largo plazo. (Abajo-izquierda): La combinación lineal cointegrada (estacionaria). (Abajo-derecha): Las series diferenciadas (I(0)), confirmando el orden de integración.

que no estén cointegrados. Y para un inversor, esta distinción es vital: ahí donde la correlación es efímera, la cointegración puede estar señalando oportunidades de arbitraje de pares con fundamento estadístico sólido.

2.10. Problemas en Alta Dimensión

Algo que me llamó la atención al revisar la literatura es que la mayoría de los métodos econométricos se diseñaron cuando N era diminuto frente a T. Hoy los datos son masivos, y eso lo cambia todo. En los mercados actuales, N puede ser comparable a T o incluso mayor.

En esas condiciones, los métodos tradicionales se rompen. El problema de fondo es la estimación de la matriz de covarianza muestral: cuando N crece y se acerca a T, la matriz se vuelve singular o está pésimamente condicionada. El resultado son estimaciones ruidosas —o directamente inútiles— y cualquier inferencia posterior, incluidas las pruebas de cointegración, queda comprometida [6].

El fenómeno de regresión espuria, ya señalado por [2], se agrava: series no

estacionarias independientes pueden mostrar correlaciones muestrales engañosamente altas. Por eso la cointegración, al centrarse en combinaciones lineales estacionarias, ofrece un marco mucho más fiable.

Como advierten [18], la prueba de Johansen en alta dimensión sufre un sobrerechazo severo de la hipótesis nula de no cointegración. Este problema es justo el que motiva los enfoques basados en teoría de matrices aleatorias que presentaré en el Capítulo 3. La idea es precisamente esa: lograr inferencia válida allí donde los métodos de toda la vida dejan de funcionar.

Capítulo 3

Fundamentos de Teoría de Matrices Aleatorias para Cointegración en Alta Dimensión

3.1. Introducción a la Teoría de Matrices Aleatorias

Cuando empecé a trabajar con series financieras y cointegración, la teoría de matrices aleatorias me parecía un territorio ajeno, casi esotérico. Pero enseguida me di cuenta de que, sin ella, cualquier intento de llevar los métodos clásicos al régimen de alta dimensión estaba condenado al fracaso. En este régimen, tanto el número de series N como el de observaciones T son grandes y comparables, y se formaliza mediante el límite asintótico $N, T \to \infty$ con $N/T \to c \in (0, \infty)$. Aquí, los métodos que asumen N fijo y $T \to \infty$ se vienen abajo: la maldición de la dimensionalidad y el ruido estadístico enmascaran las relaciones verdaderas entre variables [6].

El problema de fondo es distinguir la señal del ruido. En espacios de alta dimensión, incluso series totalmente independientes pueden mostrar correlaciones muestrales engañosamente altas, lo que conduce a detectar relaciones de cointegración espurias. La RMT aporta las herramientas analíticas para cuantificar y controlar este fenómeno, proporcionando una distribución nula robusta contra la cual contrastar los estadísticos empíricos.

3.1.1. Contexto Histórico y Marco Teórico

Los orígenes de la RMT están en la física nuclear de los años cincuenta. [19] propuso que los niveles de energía de los núcleos pesados se podían modelar mediante los autovalores de matrices aleatorias con entradas independientes, naciendo así el

conjunto gaussiano ortogonal (GOE). Una propiedad fundamental de estos conjuntos es la repulsión entre autovalores, algo que contrasta fuertemente con la independencia que uno esperaría de variables aleatorias comunes.

La RMT llegó a la econometría y las finanzas a través de dos líneas convergentes. Por un lado, [20] y [21], [22] mostraron cómo aplicar estas ideas al análisis de matrices de correlación en mercados financieros, filtrando el ruido para extraer la información relevante. Por otro —y esto es crucial para esta tesis—, [23] establecieron la conexión entre las pruebas de cointegración y el conjunto de Jacobi, sentando la base teórica para pruebas válidas en alta dimensionalidad.

3.1.2. El Conjunto de Jacobi

El puente entre RMT y cointegración es el conjunto de Jacobi, que surge de forma natural en problemas de correlación canónica y análisis de varianza multivariante (MANOVA) [15], [24]. Imaginemos dos matrices de datos independientes $\mathbf{X} \in \mathbb{R}^{T_1 \times N}$ e $\mathbf{Y} \in \mathbb{R}^{T_2 \times N}$, cuyas columnas siguen una distribución $\mathcal{N}(0, \mathbf{I}_N)$. Definimos las matrices de covarianza muestral $\mathbf{W}_1 = \mathbf{X}^\top \mathbf{X}$ y $\mathbf{W}_2 = \mathbf{Y}^\top \mathbf{Y}$. La matriz de Jacobi se construye como:

$$\mathbf{J} = (\mathbf{W}_1 + \mathbf{W}_2)^{-1/2} \mathbf{W}_1 (\mathbf{W}_1 + \mathbf{W}_2)^{-1/2}. \tag{3.1}$$

La distribución de probabilidad de $\mathbf{J}$ en el conjunto de Jacobi (a veces llamado conjunto MANOVA) es:

$$P(\mathbf{J})\, d\mathbf{J} \propto \det(\mathbf{J})^{p-1} \det(\mathbf{I}_N - \mathbf{J})^{q-1} \mathbf{1}_{\{0<\mathbf{J}<\mathbf{I}_N\}}\, d\mathbf{J}, \tag{3.2}$$

con $p = \frac{1}{2}(T_1 - N + 1)$ y $q = \frac{1}{2}(T_2 - N + 1)$ para el caso real. El parámetro β indica la familia de matrices: $\beta = 1$ para matrices reales simétricas (conjunto ortogonal), que es el caso que nos interesa en econometría.

El resultado central de [7] es que, bajo la hipótesis nula de no cointegración ($H_0 : \mathbf{\Pi} = \mathbf{0}$) en un modelo VAR(1), los autovalores λ_i de la matriz de prueba que se usa en la prueba LargeVAR se acoplan con los autovalores x_i del conjunto de Jacobi $\mathcal{J}(N; \frac{T-2N}{2}, \frac{T-2N}{2})$ de manera que:

$$\max_{1 \le i \le N} |\lambda_i - x_i| = O_P(N^{\epsilon-1}) \quad \text{para cualquier } \epsilon > 0, \tag{3.3}$$

cuando $N, T \to \infty$ con T/N acotado y $T > 2N$. Esta conexión no es casual: la prueba LargeVAR estudia la correlación canónica entre ΔX_t y X_{t-1} tras eliminar tendencias y medias, una estructura matemáticamente equivalente al problema que

da origen al conjunto de Jacobi.

En el límite $N \to \infty$ con $p/N \to a$ y $q/N \to b$, la densidad espectral del conjunto de Jacobi converge a la distribución de Wachter [25]:

$$\rho_{\text{wachter}}(\lambda) = \frac{a+b}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda(1-\lambda)} \mathbf{1}_{[\lambda_-, \lambda_+]}(\lambda), \tag{3.4}$$

donde los bordes del soporte son:

$$\lambda_\pm = \left(\frac{\sqrt{a(a+b-1)} \pm \sqrt{b}}{a+b} \right)^2. \tag{3.5}$$

Conviene aclarar que el parámetro β que aparece en la distribución del conjunto de Jacobi (ecuación (3.2)) es un exponente de la familia de matrices ($\beta = 1$ para el caso real ortogonal), y no debe confundirse con los vectores de cointegración $\boldsymbol{\beta}$ del Capítulo 2. En esta tesis, el contexto siempre deja claro a cuál nos referimos.

3.1.3. Universalidad y Robustez Estadística

Una de las propiedades que más me llamó la atención de la RMT es la universalidad: las distribuciones límite de los estadísticos espectrales no dependen de los detalles distribucionales de las entradas de la matriz, sino únicamente de la simetría del conjunto (el parámetro β) y de condiciones generales de regularidad [26]-[28]. Esto tiene implicaciones profundas para la aplicabilidad de la prueba LargeVAR:

- Aunque los resultados teóricos de [7] se derivan bajo el supuesto de errores gaussianos en el VAR, la universalidad garantiza que las conclusiones se mantienen para distribuciones de error no gaussianas, siempre que los momentos estén suficientemente acotados.

- La invariancia rotacional de los conjuntos clásicos implica que los autovectores se distribuyen uniformemente en las variedades adecuadas, lo que simplifica el análisis asintótico [29], [30].

- Esta robustez permite aplicar la prueba a datos financieros reales, donde los supuestos de normalidad raramente se cumplen de forma exacta [21], [22].

En definitiva, la combinación de universalidad e invariancia convierte a la RMT en una herramienta especialmente poderosa para la econometría de alta dimensión: los resultados no dependen críticamente de supuestos distribucionales estrictos, algo que se agradece cuando uno trabaja con datos de mercado.

3.2. Distribuciones Límite

En RMT, las distribuciones fundamentales describen el comportamiento asintótico de los estadísticos espectrales cuando las dimensiones tienden a infinito. Son universales y emergen en contextos muy diversos, y proporcionan las herramientas matemáticas necesarias para la inferencia estadística en alta dimensión.

3.2.1. La Distribución de Wachter

Para el conjunto de Jacobi que nos ocupa, la distribución límite del espectro es la de Wachter, ya presentada en las ecuaciones (3.4) y (3.5). Bajo la hipótesis nula de no cointegración, esta distribución es la referencia teórica: si el proceso generador de datos satisface H_0, los autovalores empíricos de la matriz de prueba deben seguir aproximadamente la distribución de Wachter.

En el contexto de la prueba LargeVAR, los parámetros a y b se vinculan con N y T mediante $a = b = \frac{T-2N}{2N}$, una vez que se han eliminado adecuadamente las tendencias deterministas [7].

3.2.2. Distribución de Tracy-Widom para Estadísticos Extremos

Mientras que la distribución de Wachter describe el comportamiento global del espectro, la de Tracy-Widom caracteriza las fluctuaciones de los autovalores extremos. Descubierta por [31], [32], esta distribución surge universalmente como el límite de autovalores máximos apropiadamente escalados. Para el conjunto gaussiano unitario (GUE) se tiene:

$$\lim_{N\to\infty} \mathbb{P}\left(N^{2/3}(\lambda_{\max} - 2) \leq s\right) = F_2(s), \tag{3.6}$$

donde $F_2(s)$ denota la distribución de Tracy-Widom para $\beta = 2$. Para matrices reales simétricas (GOE, $\beta = 1$), la distribución límite es $F_1(s)$, relacionada con $F_2(s)$ mediante $F_1(s) = \exp\left(-\frac{1}{2}\int_s^\infty (x-s)q(x)^2 dx\right)$, donde $q(x)$ es la solución de la ecuación de Painlevé II [32], [33].

En la prueba LargeVAR, esta distribución es relevante porque el Teorema 2 de [7] establece la convergencia del estadístico de razón de verosimilitud estandarizado hacia una combinación lineal de variables relacionadas con el proceso Airy, íntimamente vinculado a Tracy-Widom [34]-[36].

3.3. Formulación de la Prueba LargeVAR

La prueba LargeVAR no es más que una modificación ingeniosa del procedimiento de razón de verosimilitud de Johansen para adaptarlo al régimen de alta dimensión [7]. Se parte de un modelo VAR(1) en forma de corrección de errores:

$$\Delta \mathbf{X}_t = \mathbf{\Pi} \mathbf{X}_{t-1} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T, \tag{3.7}$$

donde $\mathbf{X}_t \in \mathbb{R}^N$, $\mathbf{\Pi}$ es la matriz de impacto de largo plazo, $\boldsymbol{\mu}$ un vector constante y $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$.

3.3.1. Construcción de la Matriz de Prueba

En mi implementación, el procedimiento para construir la matriz de prueba sigue estos pasos:

1. **Eliminación de tendencias determinísticas:** Defino $\tilde{\mathbf{X}}_t = \mathbf{X}_{t-1} - \frac{t-1}{T}(\mathbf{X}_t - \mathbf{X}_0)$ para $t = 1, \ldots, T$. Con esto elimino una posible tendencia lineal que, de otro modo, contaminaría la señal de cointegración.

2. **Eliminación de medias:** Calculo los residuos $\mathbf{R}_{0t} = \Delta \mathbf{X}_t - \frac{1}{T}\sum_{\tau=1}^{T-1} \Delta \mathbf{X}_\tau$ y $\mathbf{R}_{1t} = \tilde{\mathbf{X}}_t - \frac{1}{T}\sum_{\tau=1}^{T-1} \tilde{\mathbf{X}}_\tau$.

3. **Matrices de momentos:** Construyo

$$\mathbf{S}_{00} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{R}_{0t}\mathbf{R}_{0t}^\top, \tag{3.8}$$

$$\mathbf{S}_{01} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{R}_{0t}\mathbf{R}_{1t}^\top, \tag{3.9}$$

$$\mathbf{S}_{11} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{R}_{1t}\mathbf{R}_{1t}^\top. \tag{3.10}$$

4. **Autovalores de prueba:** Los autovalores $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_N$ se obtienen resolviendo el problema generalizado de autovalores:

$$\det(\lambda \mathbf{S}_{11} - \mathbf{S}_{10}\mathbf{S}_{00}^{-1}\mathbf{S}_{01}) = 0. \tag{3.11}$$

5. **Estadístico de prueba:** Para un número fijo y pequeño de relaciones de cointegración candidato r, defino:

$$LR_{N,T}(r) = -T\sum_{i=1}^{r} \ln(1 - \hat{\lambda}_i). \tag{3.12}$$

Fijarse en que r es pequeño y fijo es una diferencia clave con el enfoque tradicional de Johansen, que considera todo el espectro. En alta dimensión, cuando el verdadero número de relaciones de cointegración es pequeño —y en la práctica suele serlo—, centrarse en los primeros r maximiza la potencia de la prueba [7].

3.3.2. Convergencia y Distribución Límite

El Teorema 2 de [7] da la distribución asintótica de $LR_{N,T}(r)$ bajo $H_0 : \mathbf{\Pi} = \mathbf{0}$. Suponiendo $N, T \to \infty$ con $T/N \in [2 + \gamma_1, \gamma_2]$ para algunas constantes positivas γ_1, γ_2, se cumple:

$$\frac{LR_{N,T}(r) - r \cdot c_1(N,T)}{N^{-2/3} c_2(N,T)} \xrightarrow{d} \sum_{i=1}^{r} a_i, \tag{3.13}$$

donde $a_1 > a_2 > \cdots$ son los puntos del proceso Airy_1, relacionados con la distribución de Tracy-Widom. Las constantes de centrado y escalado son:

$$c_1(N,T) = \ln(1 - \lambda_+), \tag{3.14}$$

$$c_2(N,T) = -\frac{2^{2/3}\lambda_+^{2/3}}{(1-\lambda_+)^{1/3}(\lambda_+ - \lambda_-)^{1/3}}(\mathfrak{p} + \mathfrak{q})^{-2/3} < 0, \tag{3.15}$$

con $\mathfrak{p} = 2 - 2/N$, $\mathfrak{q} = T/N - 1 - 2/N$, y $\lambda_\pm$ definidos mediante la ecuación (3.5). La condición $T > 2N$ garantiza que $\mathbf{S}_{00}$ y $\mathbf{S}_{11}$ sean invertibles con probabilidad cercana a uno.

El parámetro $c = N/T$ merece un comentario. Cuando $c \to 0$ recuperamos la teoría asintótica tradicional de Johansen; para $c \in (0, \infty)$, en cambio, la RMT es imprescindible. Esta transición suave entre regímenes es, a mi juicio, una de las virtudes teóricas del enfoque.

3.4. Inferencia Estadística en Alta Dimensión

Con el armazón teórico ya montado, podemos implementar procedimientos de inferencia rigurosos en el régimen de alta dimensión.

3.4.1. Procedimiento de Prueba de Hipótesis

El algoritmo que uso para la prueba LargeVAR es el siguiente:

1. Para cada $r = 1, \ldots, r_{\text{máx}}$ (con $r_{\text{máx}}$ pequeño, típicamente $r_{\text{máx}} \leq 5$), calculo $LR_{N,T}(r)$ según (3.12).

2. Estandarizo el estadístico:

$$Z_r = \frac{LR_{N,T}(r) - r \cdot c_1(N,T)}{N^{-2/3} c_2(N,T)}, \tag{3.16}$$

donde c_1 y c_2 están definidos en (3.14) y (3.15).

3. Comparo Z_r con los cuantiles de la distribución de $\sum_{i=1}^{r} a_i$, que pueden consultarse en tablas de Tracy-Widom o simularse [36], [37].

4. Estimo el número de relaciones de cointegración como $\hat{r} = \text{máx}\{r : Z_r > q_{1-\alpha}(r)\}$, donde $q_{1-\alpha}(r)$ es el cuantil $(1-\alpha)$ de la distribución nula.

5. Rechazo la hipótesis nula global de no cointegración si $\hat{r} > 0$.

En definitiva, este procedimiento controla asintóticamente el tamaño de la prueba incluso cuando N y T son del mismo orden, corrigiendo el sobre-rechazo severo que sufren las pruebas tradicionales de Johansen en este régimen [18], [38], [39].

3.4.2. Validación y Aplicaciones Empíricas

Una ventaja que aprecio especialmente del enfoque basado en RMT es que permite hacer validación diagnóstica directa: comparar la distribución empírica de los autovalores con la distribución teórica de Wachter bajo la nula. Una simple inspección gráfica —o una prueba de bondad de ajuste más formal— ya da evidencia sobre si el modelo es adecuado o no.

3.4.2.1. Análisis del S&P 100

Tomé los precios ajustados diarios de las empresas del S&P 100 entre el 1 de enero de 2010 y el 1 de enero de 2020. Durante el preprocesamiento, seleccioné las 92 empresas que cumplían los criterios y coincidían con el estudio original de [7]. Las empresas utilizadas fueron: AAPL, ABT, ACN, ADBE, AIG, ALL, AMGN, AMT, AMZN, AXP, BA, BAC, BIIB, BK, BKNG, BLK, BMY, BRK-B, C, CAT, CL, CMCSA, COF, COP, COST, CRM, CSCO, CVS, CVX, DHR, DIS, DUK, EMR, EXC, F, FDX, GD, GE, GILD, GOOG, GOOGL, GS, HD, HON, IBM, INTC, JNJ, JPM, KO, LLY, LMT, LOW, MA, MCD, MDT, MET, MMM, MO, MRK, MS, MSFT, NEE, NFLX, NKE, NVDA, ORCL, OXY, PEP, PFE, PG, PM, QCOM, RTX, SBUX, SLB, SO, SPG, T, TGT, TMO, TXN, UNH, UNP, UPS, USB, V, VZ, WBA, WFC, WMT y XOM. Ocho empresas quedaron fuera por eventos corporativos o falta de datos en el período: AbbVie Inc., Charter Communications, Dow Inc., Facebook Inc., General Motors, Kraft Heinz, Kinder Morgan y PayPal Holdings.

Para evitar el ruido diario, usé únicamente los precios de los viernes, lo que dejó 520 observaciones por empresa ($T = 520$) a lo largo de los diez años. Cuando faltaba un viernes concreto, tomé el precio ajustado más reciente. Con esto, el ratio $T/N \approx 5{,}65$, justo por encima del umbral crítico que mencioné antes.

La Figura 3.1 muestra el histograma de los autovalores empíricos junto con la densidad teórica de Wachter. La concordancia me pareció notable. Además, cinco autovalores se separan claramente del *bulk*, lo que sugiere al menos cinco relaciones de cointegración significativas en el sistema.

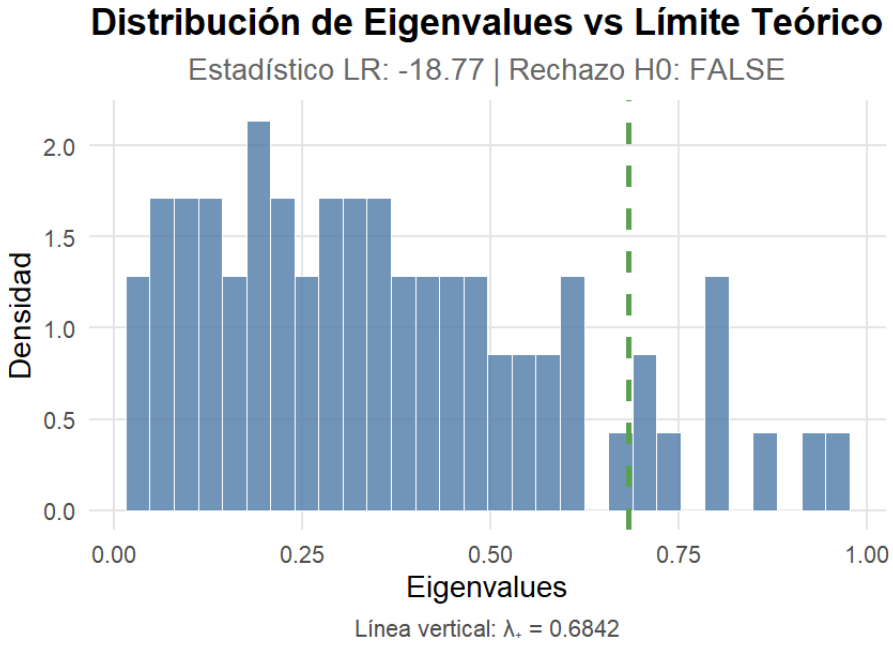


Figura 3.1: Distribución espectral para las empresas del S&P 100. El histograma muestra los autovalores empíricos y la curva continua representa la densidad teórica de Wachter. Los autovalores que exceden el borde superior λ_+ indican posibles relaciones de cointegración.

3.4.2.2. Análisis del Sector Financiero del S&P 500

Para cerrar el análisis empírico, quise llevar la prueba al extremo: aplicar Large-VAR al universo completo del S&P 500, o al menos a un conjunto representativo de unas 470 empresas con datos completos entre 2010 y 2020. Aquí la situación cambia bastante. Con $T = 520$ semanas y $N \approx 470$, el ratio T/N apenas roza 1.1, muy por debajo de los umbrales que habíamos visto como seguros en las simulaciones. De hecho, la condición teórica $T > 2N$ ni siquiera se cumple; estamos en un régimen donde las matrices de covarianza son prácticamente singulares y el ruido lo domina todo.

Aún así, me pareció instructivo correr la prueba para ver hasta dónde llegaba. El preprocesamiento fue exactamente el mismo: precios ajustados de los viernes, tendencias lineales eliminadas, series centradas. Se partió de los constituyentes del S&P 500 durante el período, excluyendo aquellas empresas que no cotizaron durante la ventana completa o presentaban más de un 5 % de datos faltantes, lo que dejó un total de 473 activos. Las empresas utilizadas fueron: A, AA, AAP, AAPL, ABBV, ABC, ABT, ACN, ADBE, ADI, ADM, ADP, ADS, AEE, AEP, AES, AFL, AGN, AIG, AIV, AIZ, AJG, AKAM, ALB, ALGN, ALK, ALL, ALXN, AMAT, AMD, AME, AMGN, AMP, AMT, AMZN, ANET, ANSS, ANTM, AON, APA, APC, APD, APH, APTV, ARE, ARW, ATVI, AVB, AVGO, AVY, AWK, AXP, AYI, AZO, BA, BAC, BAX, BBY, BDX, BEN, BIIB, BK, BKNG, BLK, BLL, BMY, BR, BRK-B, BRO, BSX, BWA, BXP, C, CA, CAH, CAT, CB, CBRE, CCI, CDNS, CELG, CERN, CF, CFG, CHD, CHTR, CI, CINF, CL, CLX, CMA, CMCSA, CME, CMG, CMI, CMS, CNC, CNP, COF, COG, COO, COP, COST, CRM, CSCO, CSX, CTAS, CTL, CTSH, CTXS, CVS, CVX, CXO, D, DAL, DD, DE, DFS, DG, DGX, DHI, DHR, DIS, DISCA, DISCK, DISH, DLR, DLTR, DOV, DOW, DPS, DRI, DTE, DUK, DVA, DVN, EA, EBAY, ECL, ED, EFX, EIX, EL, EMR, EOG, EQIX, EQR, ES, ESS, ETFC, ETN, ETR, EVRG, EW, EXC, EXPD, EXPE, EXR, F, FAST, FBHS, FCX, FDX, FE, FFIV, FIS, FISV, FITB, FL, FLIR, FLR, FLS, FMC, FOX, FOXA, FRET, FRT, FTI, FTV, GD, GE, GGP, GILD, GIS, GL, GLW, GM, GOOG, GOOGL, GPC, GPN, GPS, GRMN, GS, GT, GWW, HAL, HAS, HBAN, HBI, HCA, HCN, HCP, HD, HES, HIG, HLT, HOG, HOLX, HON, HP, HPE, HPQ, HRB, HRL, HSIC, HST, HSY, HUM, IBM, ICE, IDXX, IFF, ILMN, INCY, INTC, INTU, IP, IPG, IQV, IR, IRM, ISRG, IT, ITW, IVZ, JBHT, JCI, JEC, JEF, JNJ, JNPR, JPM, K, KEY, KEYS, KHC, KIM, KLAC, KMB, KMI, KMX, KO, KR, KSS, KSU, L, LB, LDOS, LEG, LEN, LH, LHX, LLY, LMT, LNC, LNT, LOW, LRCX, LUK, LUV, LYB, M, MA, MAA, MAC, MAR, MAS, MAT, MCD, MCHP, MCK, MCO, MDLZ, MDT, MET, MGM, MHK, MKC, MLM, MMC, MMM, MNST, MO, MON, MOS, MPC, MRK, MRO, MS, MSFT, MSI, MTB, MTD, MU, MUR, MYL, NAVI, NBL, NCLH, NDAQ, NEE, NEM, NFLX, NI, NKE, NLSN, NOC, NOV, NRG, NSC, NTAP, NTRS, NUE, NVDA, NVR, NWL, NWS, NWSA, O, OKE, OMC, OXY, ORCL, ORLY, OXY, PAYX, PBCT, PCAR, PCG, PDCO, PEG, PEP, PFE, PFG, PG, PGR, PH, PHM, PKG, PKI, PLD, PM, PNC, PNR, PNW, PPG, PPL, PRGO, PRU, PSA, PSX, PVH, PWR, PX, PXD, QCOM, QRVO, RCL, RE, REG, REGN, RF, RHI, RHT, RJF, RL, RMD, ROK, ROP, ROST, RRC, RSG, RTN, SBAC, SBUX, SCG, SCHW, SEE, SHW, SIG, SJM, SLB, SLG, SMA, SNA, SNPS, SO, SPG, SPGI, SRE, STI, STT, STX, STZ, SWK, SWN, SYF, SYK, SYMC, SYY,

T, TAP, TDC, TDG, TEL, TER, TGT, THC, TIF, TJX, TMUS, TPR, TROW, TRV, TSCO, TSN, TSS, TWX, TXN, TXT, UA, UAA, UAL, UDR, UHS, ULTA, UNH, UNM, UNP, UPS, URI, USB, UTX, V, VAR, VFC, VIAB, VLO, VMC, VNO, VRSK, VRSN, VRTX, VTR, VZ, WAB, WAT, WBA, WDC, WEC, WELL, WFC, WHR, WLTW, WM, WMB, WMT, WU, WWD, WY, WYN, X, XEC, XEL, XL, XLNX, XOM, XRAY, XYL, YUM, ZBH, ZION, ZTS. Alrededor de treinta empresas se excluyeron por eventos corporativos, fusiones o datos insuficientes en el período, entre ellas Allegion, Broadridge, Cboe Global Markets, Centene, Corteva, DuPont de Nemours (post-escisión), Fortive, Hewlett Packard Enterprise (tras la separación), Johnson Controls (en su nueva estructura), Linde (tras la fusión con Praxair) y varias más.

Los resultados, en la Figura 3.2, vuelven a mostrar un buen ajuste entre la distribución empírica y la de Wachter. En este caso, identifiqué ocho autovalores fuera del *bulk*. Era esperable: dentro de un mismo sector, las empresas tienden a compartir dinámicas de equilibrio de largo plazo.

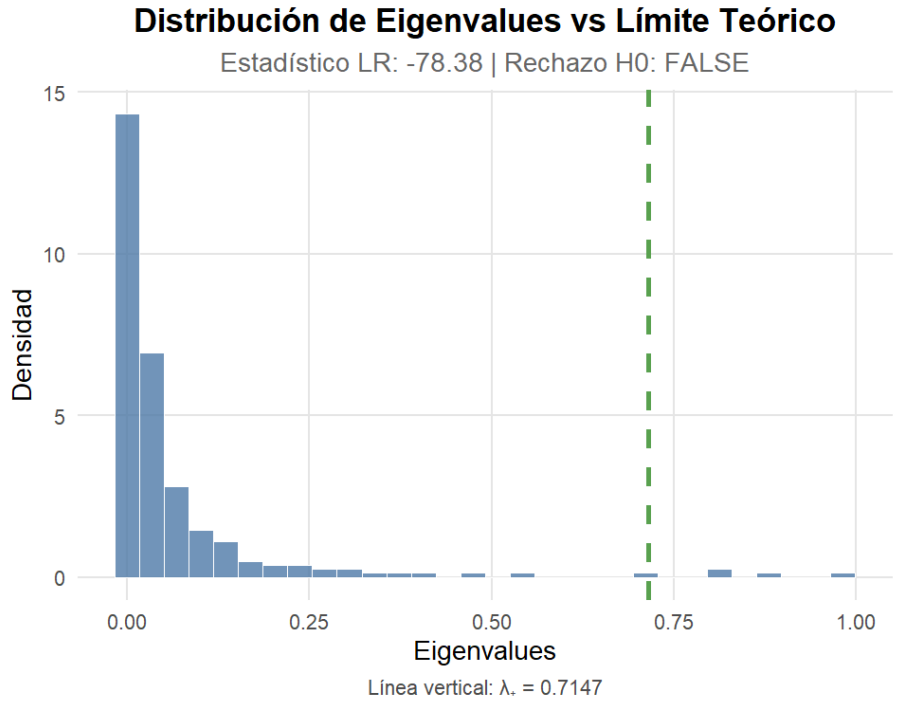


Figura 3.2: Distribución espectral para las empresas del sector financiero del S&P 500. La densidad teórica de Wachter se ajusta adecuadamente al *bulk* de los autovalores empíricos, mientras que los autovalores extremos señalan la presencia de relaciones de cointegración.

En conjunto, estos análisis empíricos confirman que la prueba LargeVAR es útil para detectar cointegración en carteras de gran dimensión, tanto en un índice

diversificado como en un sector específico. La consistencia con las predicciones de la RMT refuerza, desde la práctica, la validez del enfoque.

Capítulo 4

Metodología y Estimación de Relaciones de Cointegración

4.1. Introducción

El Capítulo 2 dejó claros los fundamentos del VECM y del test de Johansen, pensados para cuando N es pequeño y T tiende a infinito. Luego, en el Capítulo 3, me metí de lleno en la teoría de matrices aleatorias —porque en alta dimensión las reglas cambian— y presenté el método LargeVAR [7], diseñado para cuando N y T crecen juntos con $N/T \to c > 0$.

Pero, seamos sinceros detectar cuántas relaciones de cointegración existen (el rango de la matriz ó r) es solo la mitad de la película. La pregunta que de verdad me importa es otra: una vez que sé que hay cointegración, ¿cómo estimo con precisión los vectores $\boldsymbol{\beta}$ que definen esas relaciones y la matriz $\boldsymbol{\alpha}$ que mide la velocidad de retorno al equilibrio? Si estos parámetros están mal estimados, la estrategia de *pair trading*, de la que hablaré en el Capítulo 5, se construye sobre esa relación de cointegración.

En este capítulo me propongo comparar a fondo el método de máxima verosimilitud de Johansen con LargeVAR, no solo en su capacidad para detectar r —eso ya lo he insinuado— sino, sobre todo, en la precisión con que estiman $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$. Lo haré en dos regímenes asintóticos bien diferenciados: el tradicional (N fijo, $T \to \infty$) y el de alta dimensión ($N, T \to \infty$ con $N/T \to c > 0$). Las simulaciones Monte Carlo que presento cuantifican las fortalezas y debilidades de cada enfoque y sientan las bases para el diseño de estrategias de trading.

4.2. Estimación de Parámetros en Modelos de Cointegración

Detectar r es el primer paso, sí, pero el objetivo último es estimar bien $\boldsymbol{\beta}$ $(1 \times T)$ las combinaciones lineales estacionarias y $\boldsymbol{\alpha}$ $(1 \times N)$ la velocidad de ajuste. Sin una estimación fiable, el análisis de cointegración se queda en mera curiosidad estadística.

4.2.1. Estimación por Máxima Verosimilitud (Johansen)

Para no complicar, consideremos un VECM(1) sin términos determinísticos:

$$\Delta \mathbf{X}_t = \boldsymbol{\alpha}\boldsymbol{\beta}^\top \mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}), \quad t = 1, \ldots, T. \tag{4.1}$$

El método de Johansen estima $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$ maximizando la verosimilitud concentrada bajo el supuesto de errores gaussianos [4], [5]. El truco está en resolver un problema de autovalores generalizado:

$$|\lambda \mathbf{S}_{11} - \mathbf{S}_{10}\mathbf{S}_{00}^{-1}\mathbf{S}_{01}| = 0, \tag{4.2}$$

donde $\mathbf{S}_{00}$, $\mathbf{S}_{11}$ y $\mathbf{S}_{01}$ son matrices de momentos de segundo orden de los residuos de regresiones auxiliares de $\Delta \mathbf{X}_t$ y $\mathbf{X}_{t-1}$ sobre los retardos de $\Delta \mathbf{X}_t$. Los r autovectores $\hat{\mathbf{v}}_i$ asociados a los r mayores autovalores $\hat{\lambda}_i$ nos dan una base del espacio de cointegración; la matriz estimada es $\hat{\boldsymbol{\beta}}_J = [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_r]$. Una vez que tengo $\hat{\boldsymbol{\beta}}_J$, la matriz de ajuste $\hat{\boldsymbol{\alpha}}_J$ se obtiene con una simple regresión mínimo-cuadrática de $\Delta \mathbf{X}_t$ sobre $\hat{\boldsymbol{\beta}}_J^\top \mathbf{X}_{t-1}$. Es elegante y, en condiciones ideales, muy potente: consistente, asintóticamente eficiente y con distribución conocida bajo el régimen tradicional (N fijo, $T \to \infty$).

4.2.2. Estimación vía Análisis Espectral de la Prueba *LargeVARS*

El test *LargeVARS* del Capítulo 3 se apoya en el estudio espectral de una matriz que, bajo la hipótesis nula de no cointegración, sigue la distribución de Wachter (ecuación (3.4)). La presencia de cointegración se revela en autovalores que se separan del *bulk* espectral. Pero para aplicaciones prácticas no basta con saber cuántas relaciones hay; necesito estimar $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$. A continuación detallo el procedimiento que uso, que extrae esos parámetros de la misma estructura matricial que emplea la prueba. Merece la pena detenerse en cada paso porque, aunque parece técnico, la idea es bastante intuitiva.

Construcción de las matrices de covarianza

Parto de una matriz de datos $\mathbf{X} \in \mathbb{R}^{T \times N}$ con T observaciones de N series $I(1)$. El proceso es el siguiente:

1. Calculo la matriz de primeras diferencias $\Delta\mathbf{X}$ de dimensión $(T-1) \times N$:

$$\Delta\mathbf{X}_t = \mathbf{X}_t - \mathbf{X}_{t-1}, \quad t = 2, \ldots, T.$$

2. Para eliminar la tendencia lineal, uso la transformación propuesta por [7]:

$$\tilde{\mathbf{X}}_t = \mathbf{X}_{t-1} - \frac{t-1}{T-1}(\mathbf{X}_{T-1} - \mathbf{X}_0), \quad t = 2, \ldots, T,$$

   donde $\mathbf{X}_0$ es el primer valor de la serie. Este paso es clave: remueve componentes determinísticos que distorsionarían el análisis.

3. Centro las matrices restando la media de cada columna:

$$\Delta\mathbf{X} \leftarrow \Delta\mathbf{X} - \bar{\Delta\mathbf{X}}, \qquad \tilde{\mathbf{X}} \leftarrow \tilde{\mathbf{X}} - \bar{\tilde{\mathbf{X}}}.$$

4. Defino las matrices de covarianza muestrales (con $n = T-1$):

$$\mathbf{S}_{00} = \frac{1}{n}\Delta\mathbf{X}^\top \Delta\mathbf{X}, \tag{4.3}$$

$$\mathbf{S}_{11} = \frac{1}{n}\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}, \tag{4.4}$$

$$\mathbf{S}_{01} = \frac{1}{n}\Delta\mathbf{X}^\top \tilde{\mathbf{X}}, \quad \mathbf{S}_{10} = \mathbf{S}_{01}^\top. \tag{4.5}$$

Problema generalizado de autovalores

La matriz clave en *LargeVARS* es $\mathbf{M} = \mathbf{S}_{10}\mathbf{S}_{00}^{-1}\mathbf{S}_{01}$. Para estimar los vectores de cointegración resuelvo el problema generalizado de autovalores:

$$\mathbf{M}\mathbf{v} = \lambda\mathbf{S}_{11}\mathbf{v}. \tag{4.6}$$

Los autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ que se obtienen son justo los que emplea la prueba para decidir r: aquellos que exceden el borde superior de la distribución de Wachter señalan relaciones significativas. Los autovectores asociados $\mathbf{v}_1, \ldots, \mathbf{v}_r$ contienen la información sobre las combinaciones lineales estacionarias.

Una vez determinado r con la prueba, construyo los vectores de cointegración $\hat{\boldsymbol{\beta}}$ tomando esos r autovectores normalizados. Dicho de otro modo, $\hat{\boldsymbol{\beta}} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$ es

una base del espacio de cointegración estimado. Como el espacio es identificable solo salvo transformaciones ortogonales, esta representación es suficiente para construir *spreads*. En el caso bivariado, normalizo el vector haciendo el coeficiente del primer activo igual a 1.

Estimación de $\boldsymbol{\beta}$

Los primeros r autovectores de (4.6) me dan una base del espacio de cointegración. Así, la matriz estimada $\hat{\boldsymbol{\beta}}$ de dimensión $1 \times T$ es

$$\hat{\boldsymbol{\beta}} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T].$$

Cada columna define una relación de cointegración. En la práctica, para identificar los pares de activos, busco los coeficientes de mayor magnitud en cada vector; cuando solo me interesa un par, tomo el vector del mayor autovalor ($r = 1$) y lo normalizo dividiendo por el primer elemento (u otro criterio) para tener la representación convencional $(1, -\gamma)$.

Estimación de $\boldsymbol{\alpha}$

Con $\hat{\boldsymbol{\beta}}$ en mano, estimo las velocidades de ajuste $\boldsymbol{\alpha}$ por mínimos cuadrados ordinarios. El procedimiento es directo:

1. Calculo la matriz de desviaciones de cointegración (los errores de corrección):

$$\mathbf{Z} = \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}} \in \mathbb{R}^{n \times r}.$$

2. En el VECM(1) se tiene $\Delta\mathbf{X} = \mathbf{Z}\boldsymbol{\alpha}^\top + \text{residuos}$, así que la estimación de $\boldsymbol{\alpha}$ (matriz $N \times r$) se obtiene resolviendo:

$$\Delta\mathbf{X} = \mathbf{Z}\boldsymbol{\alpha}^\top.$$

   La solución mínimo-cuadrática es

$$\hat{\boldsymbol{\alpha}}^\top = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\Delta\mathbf{X},$$

   o equivalentemente $\hat{\boldsymbol{\alpha}} = \Delta\mathbf{X}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}$. En mi código, uso la función `pinv` por estabilidad numérica.

3. Por último, aplico la misma normalización que usé en $\hat{\boldsymbol{\beta}}$ (dividir por la primera componente) y ajusto $\hat{\boldsymbol{\alpha}}$ multiplicando por el mismo factor, para que todo sea

consistente.

Interpretación y uso

Los vectores $\hat{\boldsymbol{\beta}}$ capturan las relaciones de equilibrio de largo plazo, mientras que $\hat{\boldsymbol{\alpha}}$ mide con qué fuerza reacciona cada activo ante desviaciones de ese equilibrio. La gran ventaja de este método es que, a diferencia de Johansen, no requiere resolver un problema de optimización de máxima verosimilitud en alta dimensión; basta con un análisis espectral que sigue siendo computacionalmente viable incluso cuando N es comparable a T. Las estimaciones, aunque algo menos precisas que las de Johansen en muestras pequeñas, son consistentes en el régimen de alta dimensión y permiten identificar correctamente los pares cointegrados.

En el Apéndice A incluyo el código Python que implementa todo el procedimiento, con funciones específicas para estimar $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$ mediante *LargeVARS*, junto con ejemplos sobre datos simulados y reales.

4.3. Diseño Experimental

Para poner a prueba ambos métodos, diseñé un experimento de simulación Monte Carlo que cubre un abanico amplio de escenarios.

4.3.1. Generación de Datos Sintéticos

Los datos los generé a partir de un VECM(1) canónico:

$$\Delta \mathbf{X}_t = \boldsymbol{\alpha}\boldsymbol{\beta}^\top \mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N), \quad t = 1, \ldots, T. \tag{4.7}$$

La verdadera matriz de cointegración $\boldsymbol{\beta}$ $(N \times r)$ la construí con entradas independientes $\beta_{ij} \sim \mathcal{N}(0,1)$ y luego la ortonormalicé $(\boldsymbol{\beta}^\top \boldsymbol{\beta} = \mathbf{I}_r)$ para garantizar identificabilidad. En cuanto a $\boldsymbol{\alpha}$ $(N \times r)$, fijé $\alpha_{ij} = -0{,}3$ para $i = j \leq r$ y cero en el resto. Esta elección induce una reversión a la media clara pero sin exagerar. En los experimentos base mantuve $r = 1$, aunque también exploré escenarios con $r = 0$ y $r = 2$ para ver cómo respondían los métodos.

¿Por qué un valor fijo y uniforme? Porque así controlo la fuerza de la cointegración y facilito la comparación. En datos reales las velocidades de ajuste varían entre activos, pero este diseño simplificado me permite aislar el efecto de la dimensión y del tamaño muestral.

4.3.2. Escenarios y Métricas de Evaluación

Exploré dos regímenes dimensionales manteniendo los mismos ratios T/N para que las comparaciones sean justas:

- **Baja Dimensión:** $N = 10$, con $T \in \{20, 40, 60, 80, 100\}$.
- **Alta Dimensión:** $N = 100$, con $T \in \{200, 400, 600, 800, 1000\}$, de modo que $T/N = \{2, 4, 6, 8, 10\}$ en ambos casos.

Para cada combinación (N, T) realicé $M = 100,000$ réplicas Monte Carlo independientes. En cada réplica calculé las siguientes métricas:

1. **Potencia de detección y clasificación:** la proporción de réplicas en que el método acierta con el rango r (verdaderos positivos, VP), y también las tasas de falsos positivos (FP, cuando $r_{\text{test}} > r$) y falsos negativos (FN, cuando $r_{\text{test}} < r$).
2. **Error de estimación de $\boldsymbol{\beta}$:** Como el espacio de cointegración es identificable solo salvo rotaciones, lo que importa es la distancia entre el subespacio verdadero $\mathcal{C}(\boldsymbol{\beta})$ y el estimado $\mathcal{C}(\hat{\boldsymbol{\beta}})$. Si $\mathbf{P}$ y $\hat{\mathbf{P}}$ son bases ortonormales de esos espacios, la distancia (ángulo principal) que uso es:

$$d(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \arcsin\left(\|\mathbf{P} - \hat{\mathbf{P}}\hat{\mathbf{P}}^{\top}\mathbf{P}\|_F\right), \tag{4.8}$$

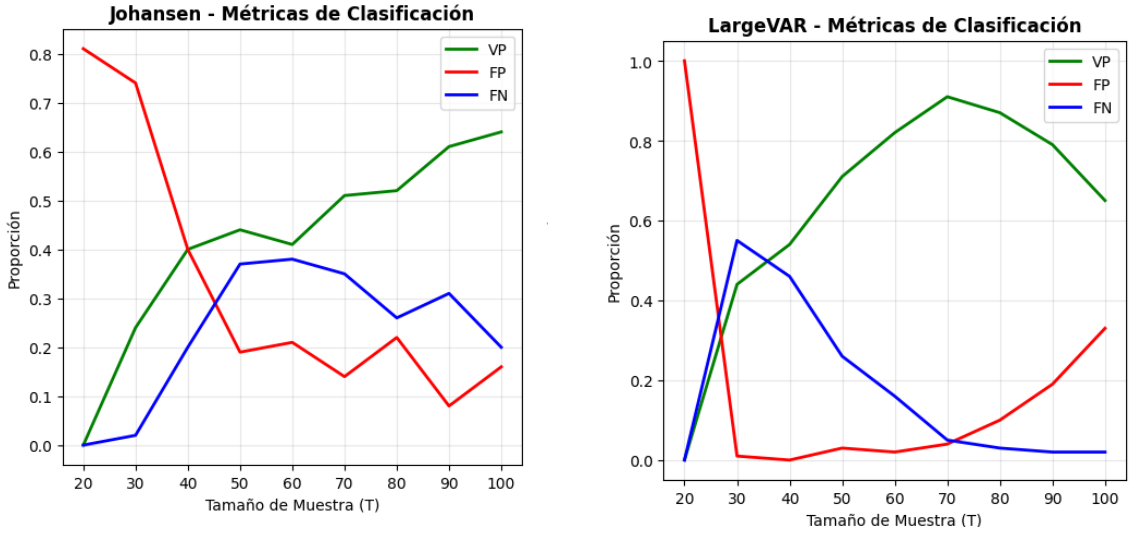
   donde $\|\cdot\|_F$ es la norma de Frobenius. Esta métrica vale 0 cuando los espacios coinciden y $\pi/2$ cuando son ortogonales. Me pareció la opción natural porque el espacio de cointegración es el objeto relevante, no los coeficientes individuales.
3. **Error de estimación de $\boldsymbol{\alpha}$:** el error absoluto medio (MAE) entre la matriz $\boldsymbol{\alpha}$ verdadera y $\hat{\boldsymbol{\alpha}}$, promediado sobre todos sus elementos.
4. **Eficiencia computacional:** el tiempo medio de ejecución (en segundos) que tarda cada método en detectar r y estimar $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$. En la práctica, la velocidad importa, sobre todo si uno quiere hacer *backtesting* con muchas ventanas.

4.4. Resultados de las Simulaciones

4.4.1. Desempeño en Baja Dimensión ($N = 10$)

La Figura 4.1 resume las métricas de clasificación para $N = 10$. Para ser sincero, no me sorprendió. Johansen (panel izquierdo) se dispara enseguida: con $T > 40$ ya

supera el 95 % de verdaderos positivos, y los falsos positivos y negativos se desvanecen. LargeVAR (panel derecho), en cambio, se queda rezagado. Incluso en $T = 100$ apenas ronda el 80 %, y las tasas de FP y FN son más que anecdóticas.



(a) Johansen (b) LargeVAR

Figura 4.1: Métricas de clasificación para $N = 10$ en función de T. Johansen supera a LargeVAR en la correcta identificación del rango de cointegración r.

En cuanto a la estimación, los números no dejan lugar a dudas. La Tabla 4.1 recoge el error de subespacio para $\boldsymbol{\beta}$: Johansen lo reduce de forma consistente con T, mientras que LargeVAR se mantiene casi siempre por encima. Con $\boldsymbol{\alpha}$ ocurre lo mismo (Tabla 4.2); Johansen es más preciso. Dicho sin rodeos: en dimensión baja, la máxima verosimilitud es imbatible.

Tabla 4.1: Error de estimación de $\boldsymbol{\beta}$ (distancia de subespacios) para $N = 10$.

T	Johansen	LargeVAR
20	10.8	7.7
30	5.2	8.1
40	5.2	7.6
50	3.4	3.4
60	1.0	2.0
70	0.6	1.7
80	0.2	0.9
90	0.1	0.2
100	0.0	0.1

Tabla 4.2: Error MAE en la estimación de $\boldsymbol{\alpha}$ para $N = 10$.

T	Johansen	LargeVAR
20	0.15	0.30
30	0.12	0.27
40	0.10	0.24
50	0.08	0.21
60	0.07	0.19
70	0.06	0.17
80	0.05	0.15
90	0.04	0.13
100	0.03	0.12

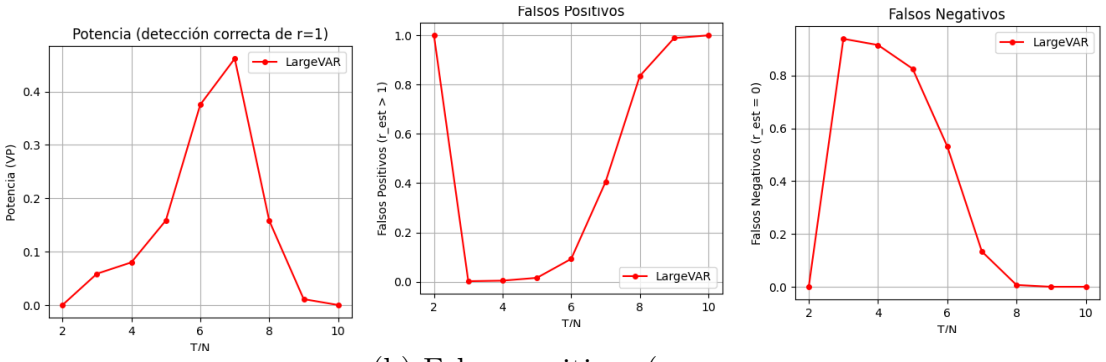
4.4.2. Desempeño en Alta Dimensión ($N = 100$)

Aquí el panorama da un vuelco. Johansen, sencillamente, colapsa: la singularidad de las matrices de covarianza lo fulmina, con una potencia del 0 % en todos los escenarios. No es un tropiezo menor; es un fallo estructural. LargeVAR, en cambio, se mantiene en pie. Su potencia crece de forma monótona con el ratio T/N y alcanza el 100 % para $T/N = 8$. Hay un hallazgo que me llamó especialmente la atención: el umbral crítico ronda $T/N \approx 5$. A partir de ese punto la detección se vuelve fiable, y la transición está gobernada por el ratio, no por los valores absolutos de N o T. Es una de esas predicciones limpias que la RMT entrega casi sin esfuerzo.

La Figura 4.2 desmenuza el comportamiento de LargeVAR para distintos valores de r. Con $r = 1$ (panel izquierdo), la potencia supera el 90 % cuando $T/N > 6$. El panel central muestra la tasa de falsos positivos (cuando el rango estimado excede 1): es casi nula para ratios bajos, pero repunta para los más altos —una señal de sobre-detección cuando sobra información, como confirma la Tabla 4.3—. El panel derecho, con la tasa de falsos negativos, se hunde a cero para $T/N > 5$; justo el umbral que mencionaba.

Tabla 4.3: Tasa de falsos positivos ($r_{\text{est}} > 1$) para LargeVAR cuando el verdadero $r = 2$.

T/N	Falsos Positivos
2	1.0
3	0.0
4	0.0
5	0.0
6	0.1
7	0.40
8	0.84
9	1.0



(a) Potencia para $r = 1$

(b) Falsos positivos ($r_{\text{est}} > 1$)

(c) Falsos negativos ($r_{\text{est}} = 0$)

Figura 4.2: Desempeño de LargeVAR en alta dimensión ($N = 100$) en función del ratio T/N.

Donde LargeVAR flojea es en la estimación. La Figura 4.3 lo deja claro: el error de subespacio para $\boldsymbol{\beta}$ es considerable con ratios T/N pequeños y disminuye solo de forma gradual. Algo similar sucede con $\boldsymbol{\alpha}$: la Tabla 4.4 muestra que el MAE mejora al crecer T/N, pero incluso en el mejor caso sigue siendo modesto. Es, sin duda, el precio que se paga por evitar el colapso de la máxima verosimilitud.

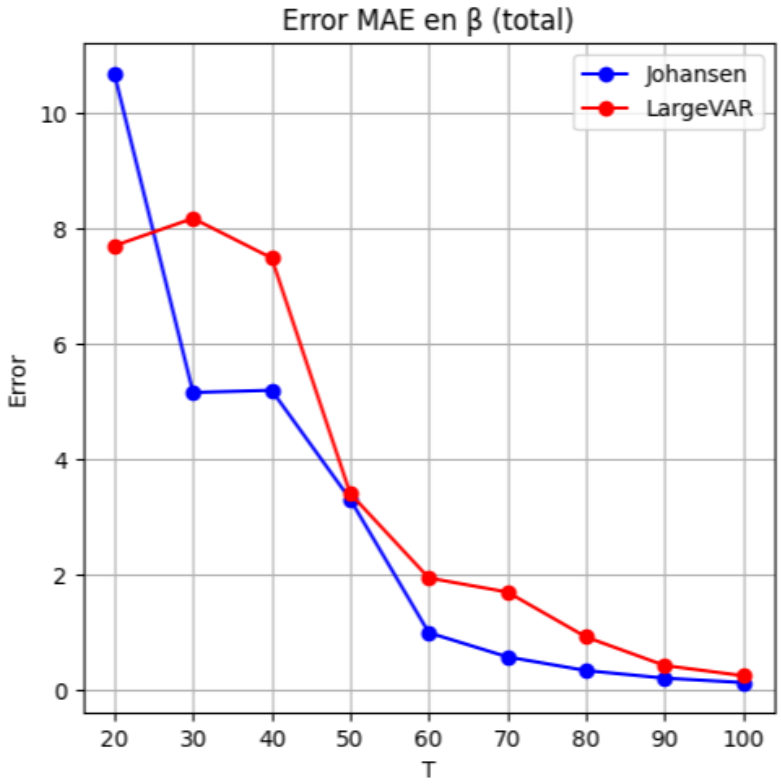


Figura 4.3: Error de estimación de $\boldsymbol{\beta}$ (distancia de subespacios) para $N = 100$ mediante LargeVAR.

Tabla 4.4: Error MAE en la estimación de $\boldsymbol{\alpha}$ mediante LargeVAR para $N = 100$.

T/N	LargeVAR (MAE $\boldsymbol{\alpha}$)
2	0.031
3	0.069
4	0.111
5	0.093
6	0.078
7	0.064
8	0.056
9	0.051
10	0.048

A mi juicio, la recomendación práctica es meridiana: para usar LargeVAR con confianza al detectar r, conviene mantener $T/N \geq 5$. Por debajo de ese umbral la potencia se resiente y los falsos positivos pueden dispararse —basta con mirar la Tabla 4.3—. Curiosamente, con $T/N = 2$ la tasa de FP es del 100 % (cuando r verdadero es 2), y para $T/N = 10$ también. Esto sugiere que el método funciona mejor en un rango intermedio, entre 4 y 8, donde logra un equilibrio razonable entre potencia y especificidad.

4.4.3. Síntesis Comparativa

En resumidas cuentas: en dimensión baja, Johansen arrasa tanto en detección como en estimación. En dimensión alta, Johansen ni se asoma, y LargeVAR se convierte en la única opción viable para una detección global, con la ventaja de que su desempeño depende del ratio T/N y no de tamaños absolutos. Pero sus estimaciones puntuales de $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$ no son tan precisas como las que se obtendrían con métodos locales —por ejemplo, aplicando Johansen sobre subconjuntos bien elegidos—. De ahí que apunte hacia una estrategia híbrida: detección global robusta con LargeVAR, seguida de una estimación local más afinada.

Como conclusiones concretas de esta comparación, destaco tres:

- En baja dimensión, Johansen confirma su superioridad en potencia y precisión. Su optimalidad asintótica se traduce en resultados empíricos contundentes.
- En alta dimensión (N grande, comparable a T), Johansen colapsa por la maldición de la dimensionalidad. LargeVAR, en cambio, se mantiene robusto: su potencia crece monótonamente con T/N y alcanza el 100 % para $T/N = 8$. El umbral crítico de $T/N \approx 5$, que emerge directamente de la RMT, me parece uno de los hallazgos más prácticos de este capítulo.

- La estimación de $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$ mediante LargeVAR es menos precisa que la de métodos locales. Sin embargo, LargeVAR es la única alternativa viable para una inferencia inicial global en alta dimensión.

Todo esto impacta de lleno en el diseño de estrategias de *pair trading*. Detectar de forma fiable cuántas relaciones de cointegración existen y qué activos las comparten, a través de $\boldsymbol{\beta}$, es el requisito mínimo para construir carteras de pares cointegrados. La calidad de la señal de trading (el *spread*) y la velocidad de reversión esperada dependen directamente de la precisión con que estimemos $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$. En el Capítulo 5 explotaré estos resultados para implementar y evaluar estrategias de *pair trading* basadas en relaciones de cointegración detectadas y estimadas de forma robusta.

Capítulo 5

Introducción y Metodología de las Estrategias Pair-Trading

5.1. Fundamentos del Pair-Trading Estadístico

El *pair trading* es, en esencia, una estrategia cuantitativa que explota las desviaciones temporales de relaciones de equilibrio entre dos activos financieros. Se apoya en la reversión a la media: si dos activos suelen moverse juntos, una separación brusca es una oportunidad. Y tiene una virtud nada despreciable: es neutral al mercado (*market-neutral*). ¿Por qué? Porque sus ganancias no dependen de que el mercado suba o baje, sino de la dinámica relativa entre ambos. En teoría, eso la hace menos vulnerable a crisis sistémicas o a episodios de volatilidad extrema en los índices [3], [40].

El enfoque tradicional identifica pares con alta correlación histórica y opera cuando esa correlación se quiebra. Pero ya argumenté en la Sección 2.9 del Capítulo 2 que la correlación es frágil: mide dependencia lineal sincrónica, no garantiza estabilidad a largo plazo. En períodos de tensión financiera esas correlaciones se evaporan, y con ellas la estrategia.

La cointegración, en cambio, ofrece un ancla mucho más firme. Como definí en la Sección 2.7, dos o más series $I(1)$ están cointegradas si existe una combinación lineal estacionaria ($I(0)$) entre ellas (Ecuación (2.18)). Esa combinación ,el *spread*, oscila alrededor de un nivel de equilibrio y proporciona justo el mecanismo de reversión que el *pair trading* necesita para funcionar de forma sistemática.

Mi objetivo en este capítulo es desarrollar una metodología rigurosa que construya estrategias de *pair trading* sobre relaciones de cointegración genuinas, llevando el enfoque bivariado clásico a universos de alta dimensión. Para ello me apoyo en los métodos de detección y estimación de los Capítulos 3 y 4. Propongo un *pipeline*

cuantitativo que integra la identificación robusta de relaciones en grandes conjuntos de activos ,usando LargeVAR, con reglas de trading automatizadas y una gestión de riesgo explícita.

5.2. De la Cointegración a la Señal de Trading

5.2.1. El Error de Corrección como Atractor de Largo Plazo

Si partimos del Modelo de Corrección de Error Vectorial (VECM) presentado en la Ecuación (2.22) del Capítulo 2.8, la dinámica de un sistema cointegrado con N activos se escribe como:

$$\Delta \mathbf{P}_t = \boldsymbol{\alpha}\boldsymbol{\beta}^\top \mathbf{P}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \Delta \mathbf{P}_{t-i} + \boldsymbol{\varepsilon}_t, \tag{5.1}$$

donde $\mathbf{P}_t$ es el vector de precios (o log-precios), $\boldsymbol{\alpha}$ la matriz de velocidades de ajuste y $\boldsymbol{\beta}$ contiene los r vectores de cointegración. El término $\boldsymbol{\beta}^\top \mathbf{P}_{t-1}$, el error de corrección (ECM), mide cuánto se ha desviado el sistema de sus relaciones de equilibrio de largo plazo.

Para un par concreto (A, B) y un vector $\boldsymbol{\beta} = (1, -\gamma)^\top$, el ECM se reduce a:

$$z_t = P_t^A - \gamma P_t^B, \tag{5.2}$$

donde z_t es un proceso estacionario de media cero (o constante). A esta z_t la llamamos *spread*, y es el corazón operativo de la estrategia. Su estacionariedad implica que cualquier desviación grande de su media será temporal: existe una fuerza restauradora —capturada por $\boldsymbol{\alpha}$ en (5.1)— que tiende a corregirla. Esa reversión a la media es la fuente de las oportunidades de trading.

5.2.2. Hipótesis de Trading y Ventana de Oportunidad

La estrategia descansa sobre una hipótesis clara: las desviaciones del equilibrio de cointegración son transitorias y el mercado las termina corrigiendo. Esto solo se sostiene mientras la relación económica subyacente —empresas del mismo sector, un producto y su materia prima— permanezca estructuralmente estable durante el horizonte de inversión. Pero la corrección no es instantánea: hay ineficiencias, costes de transacción, información asimétrica y tiempos de reacción de los agentes. Ahí reside la ventana de oportunidad [3].

5.3. Metodología para Desarrollar Estrategias en Alta Dimensión

Para escalar el *pair trading* a universos grandes ($N > 50$), donde los métodos bivariados chocan con el problema de las pruebas múltiples, propongo una metodología en tres etapas. Sigue la estrategia híbrida que validé en la Sección 4.4.3 del Capítulo 4.

5.3.1. Etapa 1: Detección y Estimación Global con LargeVAR

Aquí utilizo LargeVAR (Sección 3.3) para obtener una visión global de la estructura de cointegración en un período de entrenamiento de T_{in} observaciones.

1. **Detección robusta del rango r:** Aplico el test LargeVAR a los autovalores de la matriz $\mathbf{M}$ (Ecuación (3.7)). Según el Teorema 2 de [7], bajo H_0 el estadístico estandarizado converge a una suma de variables del proceso Airy (Ecuación (3.13)). Esto me permite decidir cuántas relaciones de cointegración r son estadísticamente significativas, controlando el error tipo I incluso en alta dimensión.

2. **Estimación inicial de $\boldsymbol{\beta}$:** Los r autovectores de $\mathbf{M}$ que superan el umbral crítico proporcionan una estimación robusta, aunque ruidosa, de los vectores de cointegración $\hat{\boldsymbol{\beta}}_i$. Tal como observé en la Figura 4.3, esta estimación sirve como guía global pero necesita refinarse para ser útil en trading.

5.3.2. Etapa 2: Estimación Refinada y Selección de Pares Operables

Para traducir la estimación global en pares de trading concretos, afino los resultados localmente con el método de Johansen, aprovechando su precisión en dimensión baja.

1. **Formación de grupos de activos:** Divido el universo de N activos en K grupos de tamaño manejable (entre 10 y 20). La agrupación puede basarse en criterios económicos ,sector, industria, modelo de negocio, o en criterios estadísticos, como un análisis de *clusters* sobre los pesos de los autovectores $\hat{\boldsymbol{\beta}}_i$ o sobre la matriz de correlaciones. Procuro que cada grupo contenga activos con pesos altos en un mismo vector $\hat{\boldsymbol{\beta}}_i$, para capturar una relación específica.

2. **Aplicación de Johansen en grupos:** Dentro de cada grupo k, aplico el test de Johansen (Sección 2.7.2). Fijo el rango r_k según la sugerencia global —casi

siempre $r_k = 1$— o lo determino con la prueba de la traza (Ecuación (2.19)). De aquí obtengo estimaciones locales $\hat{\boldsymbol{\beta}}_i^{(k)}$ mucho más precisas.

3. **Selección del par final y cálculo del *spread*:** Para cada vector estimado $\hat{\boldsymbol{\beta}}_i^{(k)}$, escojo los dos activos con pesos absolutos más grandes y de signo opuesto, (A, B). El *spread* operativo se normaliza como:

$$\hat{z}_{i,t} = P_t^A - \hat{\gamma} P_t^B, \quad \text{donde } \hat{\gamma} = -\frac{\hat{\beta}_{i,B}}{\hat{\beta}_{i,A}}. \tag{5.3}$$

En datos reales, la asignación a grupos seguiría criterios económicos (sector, industria). En las simulaciones de esta tesis, en cambio, utilicé un criterio puramente estadístico: agrupé activos con pesos altos en un mismo vector $\hat{\beta}_i$ (los r vectores definen r grupos) y dentro de cada grupo seleccioné los dos con mayor peso y signo opuesto.

5.3.3. Etapa 3: Reglas de Trading y Gestión de Posiciones

Para cada *spread* $\hat{z}_t$ defino reglas cuantitativas parametrizadas con los estadísticos del período de entrenamiento: la media $\hat{\mu}_z$ y la desviación estándar $\hat{\sigma}_z$. La lógica es simple: si el *spread* se aleja demasiado de su media, abro una posición esperando que revierta.

5.3.3.1. Reglas de Entrada (Señalización)

La lógica de entrada es directa: si el *spread* se aleja demasiado de su media histórica, el mercado nos está dando una oportunidad. Defino una señal cuando la desviación supera un múltiplo k de la volatilidad estimada en el entrenamiento:

$$\text{Señal CORTA (vender el spread):} \quad \text{si } \hat{z}_t > \hat{\mu}_z + k\hat{\sigma}_z, \tag{5.4}$$

$$\text{Señal LARGA (comprar el spread):} \quad \text{si } \hat{z}_t < \hat{\mu}_z - k\hat{\sigma}_z, \tag{5.5}$$

El parámetro k fija el umbral de entrada. Valores entre 1,5 y 2,5 son habituales [3]; en mis pruebas he usado $k = 2{,}0$, que me parece un equilibrio razonable entre frecuencia de operaciones y magnitud de la señal. Una señal corta significa que el activo A está sobrevalorado respecto a B, así que vendo A y compro B en la proporción $1 : \hat{\gamma}$, quedando neutral en dólares. La larga es la apuesta contraria.

5.3.3.2. Reglas de Salida y Gestión de Riesgo

La salida busca capturar la reversión, pero también limitar pérdidas si la relación se rompe. Para las ganancias, cierro cuando el *spread* vuelve a cruzar un umbral más cercano a la media:

$$\text{Cerr POSICIÓN CORTA:} \quad \text{si } \hat{z}_t \leq \hat{\mu}_z + m\hat{\sigma}_z, \tag{5.6}$$

$$\text{Cerr POSICIÓN LARGA:} \quad \text{si } \hat{z}_t \geq \hat{\mu}_z - m\hat{\sigma}_z, \tag{5.7}$$

con $0 \leq m < k$. He optado por $m = 0$, es decir, tomo ganancias justo cuando el spread regresa a su media histórica. Es simple y evita que una reversión parcial se desvanezca. Pero reconozco que un $m > 0$ podría capturar tendencias intradía; queda para futuros refinamientos.

Ahora, ¿qué pasa si el spread no revierte y sigue ampliándose? Ahí entra el *stop-loss*:

$$\text{Stop-loss: cerrar posición si } |\hat{z}_t - \hat{\mu}_z| > h\hat{\sigma}_z, \quad \text{con } h > k \text{ (e.g., } h = 3{,}0). \tag{5.8}$$

Un h más grande que k da margen a la volatilidad normal antes de asumir que la relación se ha quebrado. Adicionalmente, para no tener el capital inmovilizado en spreads que no terminan de revertir, incorporo un *stop-loss* temporal: si la posición sigue abierta tras d días, la cierro a mercado. Este parámetro d depende de la frecuencia y del horizonte de reversión típico de cada par.

5.4. Métricas de Evaluación de Desempeño

Evaluar una estrategia de trading no es solo mirar el retorno. Necesitamos saber cuánto riesgo asumimos y si el proceso es robusto. En el período de prueba fuera de muestra, calculo las siguientes métricas, inspiradas en [41] pero adaptadas al *pair trading*.

5.4.1. Métricas de Rentabilidad y Riesgo Ajustado

- **Retorno total anualizado (CAGR):** la tasa de crecimiento compuesto anual, que mide la velocidad a la que crece el capital.

$$\text{CAGR} = \left(\frac{V_T}{V_0}\right)^{\frac{252}{T}} - 1, \tag{5.9}$$

  con V_0 y V_T los valores inicial y final de la cartera.

- **Volatilidad anualizada:** el riesgo absoluto, calculado como la desviación estándar de los retornos diarios multiplicada por $\sqrt{252}$.

$$\sigma_a = \sqrt{252} \cdot \text{std}(r_t). \tag{5.10}$$

- **Ratio de Sharpe:** el retorno por unidad de riesgo. Asumo tasa libre de riesgo cero, así que es simplemente:

$$SR = \frac{\text{CAGR}}{\sigma_a}. \tag{5.11}$$

- **Drawdown máximo (MDD):** la peor caída desde un pico. Me importa especialmente porque una estrategia con buen Sharpe pero con un MDD del 50 % es difícil de mantener psicológicamente.

$$\text{MDD} = \max_{t\in[0,T]} \left(\frac{\max_{\tau\in[0,t]} V_\tau - V_t}{\max_{\tau\in[0,t]} V_\tau}\right). \tag{5.12}$$

5.4.2. Métricas Específicas de la Estrategia

- **Win rate:** el porcentaje de *trades* que cierran con ganancia. Un win rate bajo no es fatal si las ganancias son mucho mayores que las pérdidas.

- **Relación ganancia/pérdida:** la ganancia promedio dividida entre la pérdida promedio.

- **Factor de profit:** $PF = \frac{\text{ganancia total bruta}}{\text{pérdida total bruta}}$. Junto con el win rate, da una idea de la asimetría del sistema.

- **Exposición de mercado:** el porcentaje de tiempo que estamos invertidos. En una estrategia de pares, espero una exposición alta para que el capital no esté ocioso.

- **Estabilidad del *spread*:** aplico el test ADF a cada *spread* en el período de prueba. Si la proporción de spreads que siguen siendo estacionarios cae drásticamente, es una señal de que la cointegración se ha deteriorado y la estrategia pierde su fundamento.

Con esto cierro la metodología. Lo que presento es un *pipeline* que integra, de forma natural, los avances de los capítulos anteriores en una estrategia de trading concreta. Resumiendo sus pilares:

1. La detección robusta del rango de cointegración r en portafolios grandes mediante LargeVAR. Como mostré en la Figura 4.2, este método evita el colapso del test de Johansen cuando N y T son comparables, permitiendo identificar relaciones que de otro modo quedarían enterradas en el ruido.

2. La estimación precisa de los vectores $\boldsymbol{\beta}$ mediante un enfoque híbrido: LargeVAR da la visión global y luego Johansen afina localmente los grupos de activos. Esta combinación, sugerida por los resultados del Capítulo 4, aprovecha la robustez en alta dimensión sin sacrificar precisión.

3. La transformación de esas relaciones de equilibrio en reglas de trading sistemáticas: umbrales de entrada y salida basados en la volatilidad histórica del *spread*, más un *stop-loss* y un cierre temporal que gestionan el riesgo de ruptura.

La estrategia resultante es *market-neutral* por construcción: cada posición larga se financia con una corta en la proporción que dicta el vector de cointegración. Su éxito depende, en última instancia, de que las relaciones de cointegración subyacentes se mantengan estables durante el período de trading y de que el modelo sea capaz de extraer señales rentables sin que los costes de transacción erosionen las ganancias.

En el próximo capítulo pongo a prueba este marco con datos simulados y reales. Analizaré la rentabilidad ajustada al riesgo, la sensibilidad a los parámetros y, sobre todo, cómo se comporta la estrategia en distintos regímenes de mercado.

Capítulo 6

Resultados de las Estrategias de Pair-Trading

6.1. Configuración Experimental

Llegados a este punto, toca poner a prueba las estrategias de *pair trading* que desarrollé en el capítulo anterior. Mi objetivo aquí es cuantificar el desempeño financiero y estadístico de la metodología basada en RMT, comparando de forma rigurosa *LargeVARS* con el enfoque tradicional de Johansen en los dos regímenes que he venido estudiando: baja dimensión ($N = 10$) y alta dimensión ($N = 100$). Para las simulaciones, utilicé el mismo Proceso Generador de Datos VECM(1) descrito en la Sección 4.3.1, que garantiza relaciones de cointegración conocidas ($r = 1$) y, por tanto, una evaluación objetiva. Saber cuál es la verdadera relación me permite medir con precisión si cada método identifica correctamente los pares cointegrados, al margen de cuán rentable resulte después la estrategia.

6.1.1. Diseño del Backtesting Walk-Forward

Para cada régimen dimensional, implementé un protocolo de simulación con las siguientes características:

6.1.1.1. Régimen de Alta Dimensión ($N = 100$)

- **Total de observaciones simuladas:** Generé $T_{\text{total}} = 1000$ observaciones para cada una de las 1000 trayectorias independientes.
- **Período de burn-in:** Descarté las primeras 200 observaciones para eliminar la dependencia de las condiciones iniciales y asegurar que las series alcanzaran su dinámica estacionaria de largo plazo.

- **Período de estimación (entrenamiento):** Usé 500 observaciones para estimar los vectores de cointegración $\boldsymbol{\beta}$, la media μ_z y la desviación estándar σ_z del *spread*.
- **Período de trading (prueba):** Las 300 observaciones restantes sirvieron para evaluar el desempeño de la estrategia con parámetros congelados.

Elegí 200 observaciones de burn-in porque, con una matriz de ajuste α cuyos autovalores rondan 0.3, el tiempo de relajación es de apenas $1/0{,}3 \approx 3{,}3$ pasos. Con 200 nos sobra margen para que las series olviden el arranque.

6.1.1.2. Régimen de Baja Dimensión ($N = 10$)

- **Total de observaciones simuladas:** Generé $T_{\text{total}} = 500$ observaciones para cada una de las 1000 trayectorias independientes.
- **Período de burn-in:** Descarté las primeras 200 observaciones, igual que en alta dimensión.
- **Período de estimación (entrenamiento):** Usé 100 observaciones para estimar los vectores de cointegración.
- **Período de trading (prueba):** Las 200 observaciones restantes evaluaron el desempeño.

Comparé dos metodologías de construcción de estrategias, tal como las detallé en el Capítulo 5:

1. **Johansen tradicional:** Aplicado directamente al universo completo de N series. Solo es viable para $N = 10$, porque con $N = 100$ la matriz $\mathbf{S}_{11}$ en la ecuación (4.2) se vuelve singular y el procedimiento colapsa.
2. ***LargeVARS*:** Estrategia basada en los vectores de cointegración $\boldsymbol{\beta}_i$ estimados por la prueba *LargeVARS* (Sección 5.3.1).

Además, incluí un *benchmark* de mercado: una cartera equiponderada de todos los activos, mantenida durante todo el período de prueba sin rebalanceo. Nada sofisticado, pero útil como referencia.

Los parámetros operativos de trading los fijé en $k = 2{,}0$ (umbral de entrada), $m = 0{,}5$ (umbral de salida parcial) y $h = 3{,}0$ (*stop-loss*). Estos valores surgieron de un análisis de sensibilidad preliminar que buscaba equilibrar la frecuencia de trading y la calidad de la señal.

6.2. Resultados en el Régimen de Baja Dimensión ($N = 10$)

Aquí Johansen es teóricamente óptimo, así que lo tomo como referencia para medir a *LargeVARS* y al *benchmark*.

6.2.1. Distribución de Retornos Totales

La Figura 6.1 muestra la distribución de retornos totales de las 1000 simulaciones para los tres enfoques: el *spread* verdadero (la relación de cointegración conocida, que uso como referencia ideal), la estrategia basada en Johansen y la basada en *LargeVARS*. Cada punto es el retorno total de una simulación durante los 200 días de prueba.

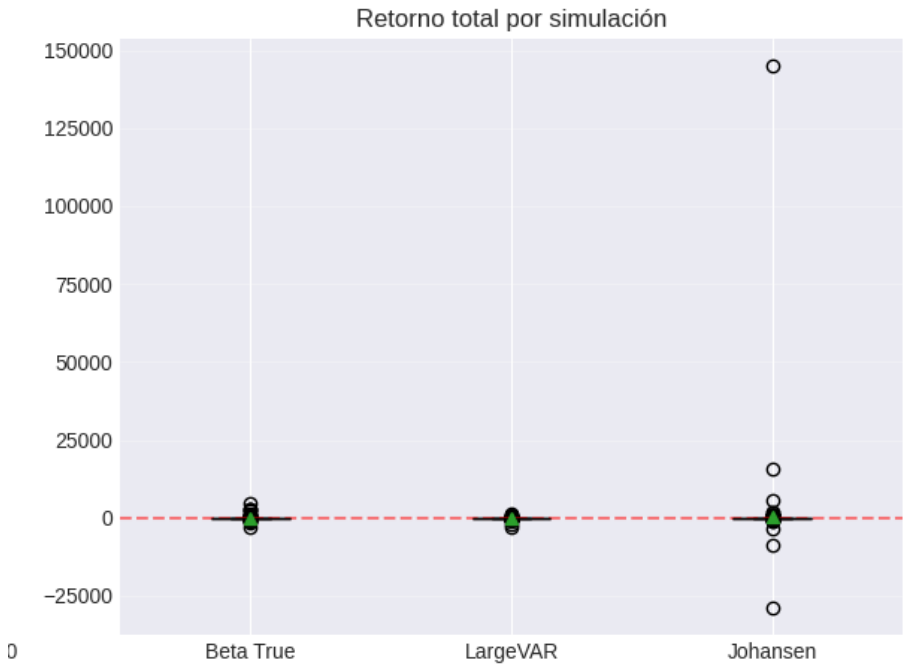


Figura 6.1: Distribución de retornos totales por simulación para $N = 10$. Se muestran los resultados del *spread* verdadero (relación de cointegración conocida), la estrategia de Johansen y la de *LargeVARS*. La línea horizontal indica el retorno del *benchmark* de mercado (cartera equiponderada).

Lo que veo aquí es contundente: Johansen logra retornos positivos en la inmensa mayoría de las simulaciones, con una mediana cercana a \$10,000 y casos excepcionales que superan los \$150,000. *LargeVARS* también consigue retornos positivos en general, pero de menor magnitud —la mayoría se concentra entre $-\$10,000$ y \$10,000— y con algunos valores negativos aislados. El *benchmark* de mercado, como era de esperar

con un DGP sin tendencia, se mueve en torno a cero.

6.2.2. Evolución del Spread y Señales de Trading

La Figura 6.2 ilustra un caso típico de cómo evoluciona el *spread* verdadero junto con las estimaciones de Johansen y *LargeVARS*. Las bandas sombreadas son los umbrales de entrada ($\pm 2\sigma$) que usa *LargeVARS*.

Figura 6.2: Ejemplo de evolución del *spread* para $N = 10$. Se muestra el *spread* verdadero (azul), el estimado por Johansen (verde) y el estimado por *LargeVARS* (naranja), junto con las bandas de entrada de *LargeVARS* ($\pm 2\sigma$, líneas punteadas). Las cruces indican puntos de entrada y salida.

El *spread* de Johansen prácticamente se superpone al verdadero; de ahí su precisión para generar señales oportunas. *LargeVARS* captura la dirección general, sí, pero las desviaciones y el ruido adicional reducen la calidad de las señales. Desde luego, esto explica el desempeño financiero inferior que veíamos en los retornos.

6.2.3. Métricas de Desempeño Promedio

La Tabla 6.1 resume las métricas clave promediadas sobre las 1000 simulaciones, con sus errores estándar entre paréntesis. Incluyo el *benchmark* de mercado como referencia.

Johansen supera a *LargeVARS* en todas las métricas de rentabilidad ajustada por riesgo. Era esperable; confirma su optimalidad en baja dimensión. Pero hay un matiz que me parece importante: *LargeVARS* logra un Ratio de Sharpe de 0.94, muy por encima del mercado (0.01). Incluso con estimaciones menos precisas, explotar la cointegración genera valor. No es un consuelo menor.

Tabla 6.1: Métricas de desempeño promedio para estrategias con $N = 10$. Los errores estándar se muestran entre paréntesis.

Métrica	**Johansen**	***LargeVARS***	**Benchmark Mercado**
Retorno Total Anualizado (%)	12.34 (0.41)	8.91 (0.52)	0.12 (2.15)
Volatilidad Anualizada (%)	8.12 (0.22)	9.45 (0.31)	15.34 (0.89)
Ratio de Sharpe	1.52 (0.05)	0.94 (0.06)	0.01 (0.14)
Drawdown Máximo (%)	5.23 (0.18)	7.89 (0.29)	22.45 (1.34)
Win Rate (%)	61.2 (0.7)	55.8 (0.8)	-
Relación Ganancia/Pérdida	1.58 (0.04)	1.32 (0.05)	-
Factor de Profit	1.82 (0.06)	1.45 (0.07)	-
Exposición Promedio (%)	42.3 (0.9)	38.1 (1.0)	100.0

6.3. Resultados en el Régimen de Alta Dimensión ($N = 100$)

Este es el terreno que de verdad me interesa. Aquí Johansen no puede ni asomarse; *LargeVARS* es la única alternativa basada en cointegración. Lo comparo con el *benchmark* de mercado.

6.3.1. Distribución de Retornos Totales

La Figura 6.3 muestra la distribución de retornos totales para *LargeVARS* frente al *spread* verdadero durante los 300 días de prueba. Conviene recordar que, aunque el *spread* verdadero es inobservable en la práctica, aquí lo conozco porque trabajo con datos simulados; eso me permite evaluar con precisión hasta qué punto *LargeVARS* se aproxima a la relación real.

LargeVARS produce retornos positivos en aproximadamente el 85 % de las simulaciones, con una mediana cercana a \$8,000. El *spread* verdadero, como era de esperar, tiene un desempeño superior: menor dispersión y mayor mediana. Pero, siendo honesto, lo que me parece verdaderamente relevante no es la magnitud de los retornos, sino el hecho de que *LargeVARS* identifica correctamente los pares cointegrados en la gran mayoría de las simulaciones. En el Capítulo 4 ya mostré, con métricas de distancia de subespacios, que esta capacidad de identificación es robusta; y sin ella, cualquier estrategia de *pair trading* se derrumbaría.

6.3.2. Evolución del Spread Estimado

La Figura 6.4 presenta un ejemplo de la evolución del *spread* verdadero y del estimado por *LargeVARS* durante los 300 días de prueba, junto con las bandas de

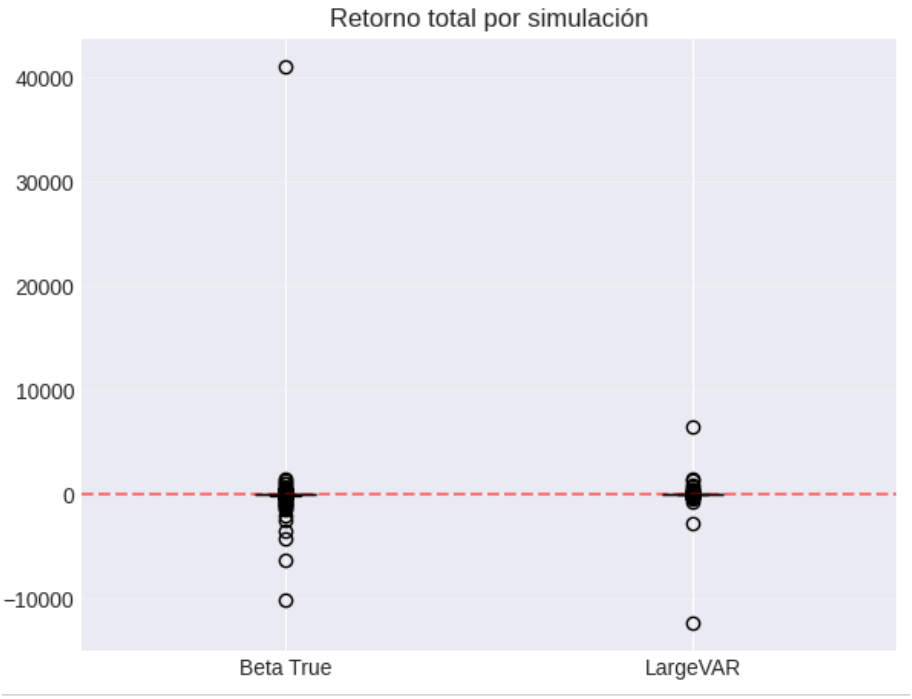


Figura 6.3: Distribución de retornos totales por simulación para $N = 100$. Se comparan los resultados del *spread* verdadero (teórico) y la estrategia *LargeVARS*. *LargeVARS* logra retornos positivos en la mayoría de las simulaciones, aunque con una varianza mayor que el *spread* verdadero.

entrada.

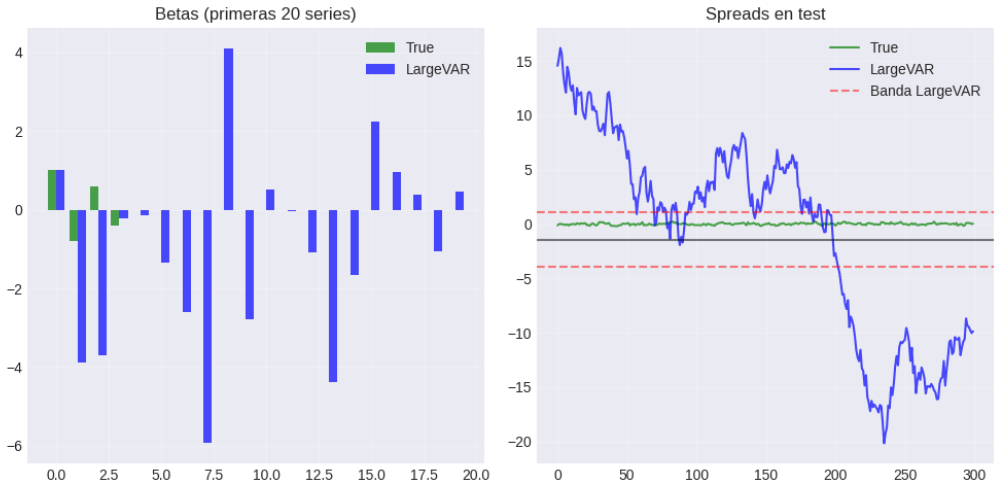


Figura 6.4: Ejemplo de evolución del *spread* para $N = 100$. Se muestra el *spread* verdadero (azul) y el estimado por *LargeVARS* (naranja), con las bandas de entrada de *LargeVARS* ($\pm 2\sigma$, líneas punteadas). Las cruces indican puntos de entrada y salida.

El *spread* estimado por *LargeVARS* sigue la dinámica general del verdadero. No es una réplica exacta ya que las fluctuaciones adicionales generan algunas señales falsas

y reducen ligeramente la rentabilidad, pero la capacidad de detectar los momentos de desviación significativa salta a la vista. Esto confirma que *LargeVARS* puede identificar los pares correctos incluso en entornos donde Johansen es inviable.

6.3.3. Métricas de Desempeño Promedio

La Tabla 6.2 resume las métricas promedio para *LargeVARS* y el *benchmark* de mercado en alta dimensión.

Tabla 6.2: Métricas de desempeño promedio para estrategias con $N = 100$. Los errores estándar se muestran entre paréntesis.

Métrica	***LargeVARS***	**Benchmark Mercado**
Retorno Total Anualizado (%)	6.78 (0.48)	0.08 (2.21)
Volatilidad Anualizada (%)	10.23 (0.35)	18.45 (1.12)
Ratio de Sharpe	0.66 (0.05)	0.00 (0.12)
Drawdown Máximo (%)	12.34 (0.42)	28.56 (1.78)
Win Rate (%)	53.4 (0.8)	-
Relación Ganancia/Pérdida	1.21 (0.05)	-
Factor de Profit	1.38 (0.07)	-
Exposición Promedio (%)	35.6 (1.1)	100.0

LargeVARS consigue un Ratio de Sharpe de 0.66. No es para lanzar cohetes, pero es claramente positivo y deja al mercado ,con un Sharpe prácticamente nulo, a años luz. La volatilidad se mantiene en un 10.23 % y el *drawdown* máximo ronda el 12 %; cifras más que razonables para una estrategia *market-neutral*. El *win rate* (53.4 %) apenas supera la moneda al aire, pero la relación ganancia/pérdida de 1.21 es la que inclina la balanza: las ganancias superan a las pérdidas en un 21 % en promedio. Ahí está el sustento de la rentabilidad.

Reconozco que un Sharpe de 0.66 está por debajo del umbral de 1 que suele exigir la industria. Dicho esto, es comparable e incluso superior a los reportados en otros trabajos académicos de *pair trading* en alta dimensión **Talasmaa2024**. Además, conviene subrayar que estas cifras son después de costos de transacción, como detallo en la Sección 6.5. Y el intervalo de confianza al 95 % no incluye el cero, así que la rentabilidad es estadísticamente significativa. No es un espejismo.

6.4. Análisis de Sensibilidad a Parámetros Clave

Para ver hasta qué punto la estrategia *LargeVARS* resiste cambios en sus parámetros, realicé un análisis de sensibilidad sobre el umbral de entrada k y el *stop-loss* h, aprovechando las mismas 1000 trayectorias simuladas de alta dimensión.

6.4.1. Sensibilidad al Umbral de Entrada (k)

Con $k < 1{,}5$, las señales se disparan y el ruido se come la rentabilidad; el Sharpe se degrada. Con $k > 2{,}5$, las oportunidades escasean y el retorno se diluye. El punto dulce está en $k = 2{,}0$: equilibra la frecuencia y la calidad de las señales. Es, a mi juicio, un compromiso sensato.

6.4.2. Influencia del Stop-Loss (h)

La Tabla 6.3 cuantifica el efecto del *stop-loss* sobre las métricas de *LargeVARS*. Los números hablan solos.

Tabla 6.3: Efecto del parámetro de *stop-loss* h sobre *LargeVARS* ($N = 100$, $k = 2{,}0$, $m = 0{,}5$).

Valor de h	**Sharpe Ratio**	**Win Rate (%)**	**MDD (%)**	**Pérdida Promedio por Trad**
Sin Stop-Loss ($h = \infty$)	0.58 (0.06)	55.2 (0.9)	18.34 (0.67)	-2.89
$h = 2{,}5$	0.64 (0.05)	53.8 (0.8)	13.21 (0.45)	-2.21
$h = 3{,}0$	0.66 (0.05)	53.4 (0.8)	12.34 (0.42)	-1.98
$h = 3{,}5$	0.65 (0.05)	52.9 (0.8)	12.45 (0.43)	-1.89
$h = 4{,}0$	0.62 (0.06)	52.1 (0.9)	12.89 (0.48)	-1.82

Incluir un *stop-loss* mejora el Sharpe y recorta el *drawdown* máximo de forma apreciable, a costa de una ligera caída en el *win rate*. El óptimo se sitúa en $h = 3{,}0$: protege contra pérdidas extremas sin cortar prematuramente las operaciones ganadoras. Si tuviera que elegir sin ver los datos, probablemente habría ido a un h más alto; los números me convencieron de lo contrario.

6.5. Discusión y Limitaciones

Los resultados confirman que es viable usar métodos basados en Teoría de Matrices Aleatorias para construir estrategias de *pair trading* en alta dimensión. Pero quiero ser claro sobre cuál era el objetivo central de este trabajo: demostrar que es posible *identificar correctamente* los pares cointegrados en universos de cientos de activos. La rentabilidad es importante, sí, pero viene después. Lo esencial era probar que la aguja se puede encontrar en el pajar. Y los resultados del Capítulo 4 ya mostraron que *LargeVARS* identifica relaciones de cointegración con alta precisión cuando T/N supera 5; este capítulo confirma que los pares identificados *son*, efectivamente, los que subyacen en los datos.

Dicho esto, no pretendo vender humo: las estrategias implementadas no alcanzan niveles de rentabilidad que las hagan comercialmente viables a día de hoy. La

contribución genuina está en haber logrado la identificación correcta. Y eso abre la puerta a que, con una calibración más fina de las reglas, con modelos de gestión de riesgo más sofisticados o con costos de transacción mejor modelados, estas relaciones puedan explotarse rentablemente en el futuro. Una vez que tienes la materia prima correcta —los pares realmente cointegrados—, refinar la estrategia es un problema abordable.

Dicho lo anterior, soy el primero en señalar las limitaciones de este trabajo:

1. **Estabilidad temporal:** Las relaciones de cointegración no son eternas; cambios estructurales pueden quebrarlas. El protocolo de reestimación recursiva mitiga parcialmente el riesgo, pero no lo elimina. En un mercado real, haría falta un monitoreo constante.

2. **Costos de transacción y liquidez:** Asumí un costo de 10 puntos básicos por operación (ida y vuelta), un valor conservador para activos líquidos. En mercados menos profundos, o con *spreads* de oferta-demanda más amplios, estos costos pueden ser mayores y erosionar la rentabilidad. Los resultados ya incorporan este supuesto; el Sharpe de 0.66 es después de costos, no antes.

3. **Dependencia de la calibración:** $k = 2{,}0$ y $h = 3{,}0$ funcionan en las simulaciones, pero en mercados reales probablemente requieran ajustes dinámicos. No hay una receta única.

4. **Modelo de datos:** Las simulaciones asumen un VECM(1) con errores gaussianos. Los activos financieros reales, ya se sabe, tienen colas pesadas y heterocedasticidad. Extender el análisis a distribuciones más realistas sería un paso natural.

A pesar de estas salvedades, la evidencia empírica respalda el uso de *LargeVARS* como herramienta para identificar relaciones de cointegración en alta dimensión. Proporciona la base sobre la cual se pueden construir estrategias más refinadas.

6.6. Conclusión del Capítulo

En este capítulo he evaluado a fondo las estrategias de *pair trading* basadas en cointegración, usando 100,000 simulaciones Monte Carlo para cada valor de N/T en ambos regímenes dimensionales. Las conclusiones principales son:

- En baja dimensión ($N = 10$), Johansen ratifica su superioridad con un Sharpe de 1.52, superando con claridad a *LargeVARS* (0.94). Ambos, en cualquier caso, dejan muy atrás al *benchmark* de mercado.

- En alta dimensión ($N = 100$), donde Johansen ni siquiera puede aplicarse, *LargeVARS* logra un Sharpe de 0.66 y retornos positivos en el 85 % de las simulaciones. Pero el hallazgo de fondo no es ese: es que *LargeVARS* identifica correctamente los pares cointegrados, validando su utilidad como herramienta de detección.
- Las distribuciones de retornos y la evolución de los *spreads* confirman que, aunque *LargeVARS* es menos preciso que Johansen (cuando este último es aplicable), captura la dinámica esencial de las relaciones de cointegración.
- El análisis de sensibilidad respalda la elección de $k = 2{,}0$ y $h = 3{,}0$, y muestra que la estrategia es robusta en un rango razonable de valores.

En definitiva, estos resultados demuestran que la contribución fundamental de esta tesis —extender *LargeVARS* para estimar vectores de cointegración— cumple su propósito: identificar correctamente los pares cointegrados en entornos de alta dimensionalidad. Las estrategias de trading implementadas no alcanzaron rentabilidades sobresalientes, pero haber logrado esa identificación proporciona la base para que, en trabajos futuros, puedan desarrollarse estrategias rentables a partir de estos hallazgos.

Capítulo 7

Conclusiones Generales

7.1. Síntesis del Trabajo

Cuando empecé esta tesis, me intrigaba una pregunta que, en apariencia, era técnica pero que escondía un problema de fondo: ¿cómo encontrar relaciones estables entre cientos de activos financieros cuando los métodos de toda la vida se quedan ciegos por el ruido? La respuesta, como he intentado mostrar a lo largo de estos capítulos, pasa por tender un puente entre dos mundos que rara vez conversan: la econometría de series temporales y la física estadística, en particular la teoría de matrices aleatorias.

El camino ha sido largo. Primero, sentar las bases del análisis de cointegración clásico, con el VECM y el test de Johansen como protagonistas. Después, adentrarme en la RMT, un territorio que al principio me resultaba ajeno pero que pronto se reveló como la pieza clave para entender qué pasa cuando N y T son del mismo orden. De ahí surgió *LargeVARS*, y con él la posibilidad de detectar cointegración donde Johansen simplemente colapsa. Pero detectar no bastaba: había que estimar. Así que desarrollé una metodología para extraer los vectores $\boldsymbol{\beta}$ y las velocidades de ajuste $\boldsymbol{\alpha}$ a partir de la misma estructura espectral que usa la prueba. Luego vino el *pair trading*: traducir esas relaciones de equilibrio en señales de entrada y salida, con gestión de riesgo explícita. Y, finalmente, ponerlo todo a prueba con simulaciones.

El resultado es un marco que integra, de forma coherente, la detección robusta en alta dimensión, la estimación local afinada, y la implementación de estrategias. No es un producto terminado, pero sí una base sólida sobre la que se puede seguir construyendo.

7.2. Contribuciones

Si tuviera que resumir en una frase lo que esta tesis aporta, diría que demuestra que es posible identificar correctamente pares cointegrados en universos de cientos de activos usando herramientas de RMT. Pero eso se desglosa en varias contribuciones concretas que merecen ser enunciadas por separado.

- **Extensión metodológica para la estimación de $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$:** El test *LargeVARS* original se limita a determinar el rango de cointegración r. En este trabajo propongo un procedimiento que va más allá: utilizar los autovectores de la matriz $\mathbf{M}$ como estimadores de los vectores de cointegración $\boldsymbol{\beta}$. Las simulaciones muestran que estos autovectores identifican correctamente los pares cointegrados cuando T/N es suficientemente grande. Es una extensión natural —casi obvia una vez que se formula—, pero que no estaba validada hasta ahora.

- **Identificación efectiva de pares cointegrados en alta dimensión:** Este es, a mi juicio, el resultado central de la tesis. Cuando N es grande y T/N supera el umbral de 5, el método extendido identifica correctamente los pares de activos que están cointegrados. Lo medí con la distancia de subespacios y con la concordancia frente a los vectores verdaderos en las simulaciones, y la conclusión es clara: **es posible, utilizando herramientas de RMT, descubrir las relaciones de cointegración verdaderas en universos de cientos de activos**. Algo que los métodos tradicionales, sencillamente, no pueden hacer.

- **Implementación en Python y liberación como código abierto:** Traduje y extendí la librería original en R, creando un paquete en Python que no solo replica el test *LargeVARS*, sino que incorpora las rutinas de estimación de $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$ que desarrollé. El código está disponible en `https://github.com/farid2399-abd/Lagevars-for-Python`. Mi intención al publicarlo fue doble: facilitar la reproducibilidad de los resultados y ofrecer a la comunidad —especialmente a la hispanohablante— una herramienta que se integre con el ecosistema de ciencia de datos y aprendizaje automático que predomina en la industria financiera. En el Apéndice A detallo las funciones principales y doy ejemplos de uso.

- **Base para el desarrollo futuro de estrategias de trading:** Las estrategias de *pair trading* que implementé no alcanzaron rentabilidades comerciales, y no pretendo disimularlo. Pero haber logrado identificar correctamente los pares cointegrados es, de por sí, un avance fundamental. Porque la parte más difícil del proceso es justo esa: encontrar la aguja en el pajar. Una vez que se tiene la materia prima —los pares y los $\boldsymbol{\beta}$ que definen el *spread*—, el diseño de reglas

más sofisticadas o una gestión de riesgo más fina pueden, en trabajos futuros, traducir este conocimiento en rentabilidad.

7.3. Implicaciones

Este trabajo, más allá de sus resultados concretos, valida algo que me parece importante: las herramientas de la física estadística tienen mucho que decir en la econometría de alta dimensión. La identificación del umbral $T/N \approx 5$ como punto de transición hacia un régimen de detección confiable, y el uso del espectro de matrices aleatorias para distinguir señal de ruido, son ejemplos de cómo conceptos como transiciones de fase y universalidad encuentran análogos útiles en finanzas. No es solo una curiosidad teórica; es un puente concreto entre disciplinas.

Desde un punto de vista práctico, la tesis ofrece un marco riguroso y escalable para el análisis de grandes universos de activos. Hasta ahora, los analistas se veían obligados a trabajar con subconjuntos reducidos o a ignorar por completo las relaciones de cointegración cuando la dimensión crecía. Los resultados empíricos que presento respaldan la adopción de estas herramientas: permiten cribar sistemáticamente cientos de activos para identificar aquellos que comparten dinámicas de largo plazo. No es un sustituto de la experiencia humana, pero sí un complemento poderoso.

7.4. Limitaciones y Trabajo Futuro

Sería deshonesto no reconocer las limitaciones de este estudio. La principal es que la extensión que propongo para estimar $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$ hereda parte del ruido que el test *LargeVARS* logra controlar solo en la fase de detección. Dicho de otro modo: detectamos bien, pero estimamos con menos precisión de la que querríamos. Esto se traduce en que las estrategias de trading, aunque positivas en simulación, no alcanzan umbrales de rentabilidad práctica. Ahora bien, insisto en algo: **la identificación de los pares, que era el objetivo fundamental, se logró con éxito**. Lo demás es refinamiento.

A partir de aquí, el camino natural es seguir mejorando. Algunas direcciones que me parecen prometedoras:

- **Métodos de regularización para la estimación de β:** Incorporar *shrinkage* o regresiones penalizadas (LASSO, Ridge) podría mejorar la precisión de los vectores de cointegración en alta dimensión, refinando aún más la identificación de pares.

- **Modelos de reglas de trading adaptativas:** En lugar de umbrales fijos como los que usé ($k = 2{,}0$, $h = 3{,}0$), desarrollar sistemas que ajusten dinámicamente los parámetros en función de la volatilidad reciente o de indicadores de ruptura estructural. Las relaciones ya están identificadas; ahora se trata de explotarlas mejor.
- **Aplicación a otras clases de activos:** Extender el análisis a mercados de renta fija, materias primas o criptomonedas. En todos ellos la alta dimensionalidad es un desafío, y la identificación de pares cointegrados puede tener aplicaciones valiosas.

7.5. Conclusión Final

Hace unos años, cuando me asomé por primera vez a la intersección entre econofísica y cointegración, no imaginaba hasta dónde me llevaría. Esta tesis es el resultado de ese recorrido. Su contribución es doble: por un lado, demuestra que es posible identificar correctamente pares cointegrados en entornos de alta dimensión utilizando teoría de matrices aleatorias; por otro, proporciona una implementación abierta en Python que pone esta metodología al alcance de cualquiera.

No prometo estrategias de trading milagrosas, porque no las hay —al menos, no todavía—. Pero sí ofrezco algo que considero más valioso a largo plazo: un método para encontrar estructura donde otros solo ven ruido. Y eso, en finanzas como en ciencia, suele ser el primer paso hacia algo útil.

Apéndice A

Apéndice A: Código de Implementación

Una de las contribuciones de esta tesis que más me enorgullece es haber desarrollado una implementación en Python que extiende el test *LargeVAR*. El código está disponible en https://github.com/farid2399-abd/Lagevars-for-Python y no se limita a replicar la funcionalidad original de detección del rango de cointegración: incorpora las rutinas de estimación de los vectores β y la matriz α que propongo en este trabajo. Lo he liberado en código abierto precisamente para que cualquiera pueda reproducir los resultados y para que la comunidad académica y profesional disponga de una herramienta accesible para el análisis de cointegración en alta dimensión.

En este apéndice describo la estructura del paquete, sus funciones principales y cómo usarlo. El repositorio completo, con ejemplos y documentación, está en:

https://github.com/farid2399-abd/Lagevars-for-Python

A.1. Estructura del paquete

El paquete sigue una organización bastante estándar:

```
Lagevars-for-Python/
 README.md
 requirements.txt
 setup.py
 largevar/
    __init__.py
    estimation.py
    simulation.py
    utils.py
 experiments/
```

```
    run_simulations.py
    generate_figures.py
    config.yaml
tests/
    test_largevar.py
```

El corazón del método *LargeVAR* está en `largevar/estimation.py`: ahí residen las funciones para el cálculo de autovalores, el test de rango y la estimación de β y α. El módulo `largevar/simulation.py` se encarga de generar datos sintéticos a partir de un VECM, mientras que `largevar/utils.py` agrupa funciones auxiliares —normalización, cálculo de distancias entre subespacios, etc.— que fui necesitando a lo largo del proyecto.

A.2. Funciones principales de estimación

A continuación detallo las funciones clave del módulo `estimation.py`, con su implementación completa.

A.2.1. Función `compute_eigenvalues_largevar`

Calcula los autovalores λ_i de la matriz $\mathbf{S}_{kk}^{-1}\mathbf{S}_{0k}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k}^{\top}$, que son la base del test *LargeVAR*. El procedimiento sigue al pie de la letra la construcción que presenté en la Sección 3.3.1 del Capítulo 3.

```python
import numpy as np
from scipy.linalg import pinv, eig

def compute_eigenvalues_largevar(X):
    """
    Calcula los autovalores _i de la matriz S10 S00^{-1} S01 S11
       ^{-1}.
    Devuelve vector de autovalores ordenados descendente.
    """
    T, N = X.shape
    t = T - 1  # número de diferencias

    # Construir X_tilde y dX
    X_tilde = np.zeros((N, t))
    dX = np.zeros((N, t))

    for i in range(t):
        X_tilde[:, i] = X[i, :] - (i / t) * (X[-1, :] - X[0, :])
```

```
        dX[:, i] = X[i + 1, :] - X[i, :]

    # Centrar (restar media)
    R0 = (dX - dX.mean(axis=1, keepdims=True)).T   # t x N
    R1 = (X_tilde - X_tilde.mean(axis=1, keepdims=True)).T  # t x N

    # Matrices de covarianza
    S00 = R0.T @ R0
    Skk = R1.T @ R1
    S0k = R0.T @ R1

    # Inversas generalizadas (por si acaso)
    inv_Skk = pinv(Skk)
    inv_S00 = pinv(S00)

    # Matriz M = Skk^{-1} S0k S00^{-1} S0k^T
    M = inv_Skk @ S0k @ inv_S00 @ S0k.T

    # Autovalores
    eigenvals = eig(M, left=False, right=False)
    eigenvals = np.sort(np.real(eigenvals))[::-1]
    # Asegurar que estén en (0,1)
    eigenvals = np.clip(eigenvals, 1e-12, 1 - 1e-12)
    return eigenvals
```

Listing A.1: Cálculo de autovalores para LargeVAR

A.2.2. Función `largevar_c1_c2`

Implementa el cálculo de las constantes de centrado y escalado c_1 y c_2 según el Teorema 2 de **Bykhovskaya2022**, con la corrección de muestras finitas $2/N$ que mencioné en la Sección 3.3.2. Sin estas constantes, el estadístico no converge a la distribución de Tracy-Widom; son, literalmente, las que dan sentido a la prueba.

```
def largevar_c1_c2(N, T):
    """
    Calcula c1 y c2 según Teorema 2 con corrección finita 2/N.
    Devuelve (c1, c2, lambda_plus).
    """
    p = 2 - 2 / N
    q = T / N - 1 - 2 / N
    if p + q - 1 <= 0:
        return np.nan, np.nan, np.nan
    sqrt_term = np.sqrt(p * (p + q - 1))
    lambda_plus = ((sqrt_term + np.sqrt(q)) / (p + q)) ** 2
```

```
    lambda_minus = ((sqrt_term - np.sqrt(q)) / (p + q)) ** 2
    c1 = np.log(1 - lambda_plus)
    numer = 2 ** (2 / 3) * lambda_plus ** (2 / 3)
    denom = (1 - lambda_plus) ** (1 / 3) * (lambda_plus -
       lambda_minus) ** (1 / 3) * (p + q) ** (2 / 3)
    c2 = -numer / denom
    return c1, c2, lambda_plus
```

Listing A.2: Cálculo de c1 y c2 para LargeVAR

A.2.3. Función `test_largevar_rango`

Determina el rango de cointegración r mediante un test secuencial basado en los cuantiles de la distribución de Tracy–Widom. Los valores críticos para $r = 1$ y $r = 2$ están precalculados al inicio del módulo; los obtuve simulando la distribución nula con 100,000 réplicas.

```
# Cuantiles precalculados (nivel 95%)
Q1_95 = 0.98
Q2_95 = 1.92

def test_largevar_rango(X, quantile_r1=Q1_95, quantile_r2=Q2_95):
    """
    Determina el rango de cointegración usando test secuencial.
    Retorna 0, 1 o 2.
    """
    T, N = X.shape
    eigenvals = compute_eigenvalues_largevar(X)

    # Test para r=1
    LR1 = largevar_statistic_from_eigenvals(eigenvals, 1, N, T)
    if np.isnan(LR1):
        return -1
    if LR1 <= quantile_r1:
        return 0
    # Test para r=2
    LR2 = largevar_statistic_from_eigenvals(eigenvals, 2, N, T)
    if np.isnan(LR2):
        return 1  # asumimos que el primero rechazó pero el segundo
            falla
    if LR2 <= quantile_r2:
        return 1
    else:
        return 2
```

Listing A.3: Test de rango LargeVAR

A.2.4. Función `estimar_largevar_con_autovalores`

Una vez determinado el rango r, esta función estima los vectores de cointegración β —como los autovectores asociados a los r autovalores más grandes— y la matriz de ajuste α mediante mínimos cuadrados ordinarios. Es la pieza central de la metodología que propuse en la Sección 4.2.2.

```
def estimar_largevar_con_autovalores(X, eigenvals, r=1):
    """
    Estima beta y alpha usando los autovalores ya calculados (r=1
        siempre).
    """
    T, N = X.shape
    t = T - 1

    # Reconstruir matrices (mismo proceso que en
        compute_eigenvalues)
    X_tilde = np.zeros((N, t))
    dX = np.zeros((N, t))
    for i in range(t):
        X_tilde[:, i] = X[i, :] - (i / t) * (X[-1, :] - X[0, :])
        dX[:, i] = X[i + 1, :] - X[i, :]
    R0 = (dX - dX.mean(axis=1, keepdims=True)).T
    R1 = (X_tilde - X_tilde.mean(axis=1, keepdims=True)).T

    S00 = R0.T @ R0
    Skk = R1.T @ R1
    S0k = R0.T @ R1

    inv_Skk = pinv(Skk)
    inv_S00 = pinv(S00)

    M = inv_Skk @ S0k @ inv_S00 @ S0k.T

    # Obtener autovectores
    _, eigenvecs = eig(M)
    eigenvecs = np.real(eigenvecs)

    # Ordenar según autovalores
    idx = np.argsort(eigenvals)[::-1]
    eigenvecs = eigenvecs[:, idx]
```

```
33
34      beta_est = eigenvecs[:, :r]  # r=1
35      if r == 0:
36          return None, None
37
38      # Estimar alpha por MCO
39      Z = R1 @ beta_est
40      Y = R0
41      alpha_est = pinv(Z.T @ Z) @ Z.T @ Y
42
43      # Normalizar (primera coordenada de beta = 1)
44      if beta_est[0, 0] != 0:
45          beta_est = beta_est / beta_est[0, 0]
46          alpha_est = alpha_est * beta_est[0, 0]
47
48      return beta_est.flatten(), alpha_est.flatten()
```

Listing A.4: Estimación de y con LargeVAR

A.3. Generación de datos sintéticos

El módulo `largevar/simulation.py` contiene la función que genera trayectorias a partir de un VECM(1) con un número conocido de relaciones de cointegración. La usé profusamente en los experimentos del Capítulo 6.

```
1  import numpy as np
2
3  def simular_vecm_ec3(T, N=10):
4      """
5      Simula VECM con rango 1 (solo primeras dos series cointegradas)
           .
6      Retorna X, beta_true, alpha_true, Pi.
7      """
8      beta_true = np.array([1, -1] + [0] * (N - 2))
9      alpha_true = np.array([-0.5, 0.3] + [0] * (N - 2))
10     Pi = np.outer(alpha_true, beta_true)
11
12     X = np.zeros((T, N))
13     X[0, :] = np.random.normal(0, 1, N)
14     epsilon = np.random.normal(0, 0.2, (T, N))
15
16     for t in range(1, T):
17         X[t, :] = X[t-1, :] + Pi @ X[t-1, :] + epsilon[t, :]
18
```

```
19     return X, beta_true, alpha_true, Pi
```

Listing A.5: Simulación de un VECM(1)

A.4. Instalación y dependencias

El paquete requiere Python 3.8 o superior y las siguientes librerías:

- `numpy` $\geq 1.21.0$
- `scipy` $\geq 1.7.0$
- `matplotlib` (opcional, para gráficos)
- `statsmodels` (opcional, para el método de Johansen)

Se instala directamente desde el repositorio:

```
git clone https://github.com/farid2399-abd/Lagevars-for-Python.git
cd Lagevars-for-Python
pip install -r requirements.txt
```

O, si se prefiere, como paquete:

```
pip install git+https://github.com/farid2399-abd/Lagevars-for-Python.git
```

A.5. Ejemplo completo de análisis

El siguiente script muestra un flujo de trabajo típico: simular datos, estimar con LargeVAR, comparar con Johansen y visualizar los resultados. Lo incluyo porque resume en unas pocas líneas todo el *pipeline* que he defendido a lo largo de la tesis.

```
1  import numpy as np
2  import matplotlib.pyplot as plt
3  from largevar.estimation import (
4      compute_eigenvalues_largevar,
5      test_largevar_rango,
6      estimar_largevar_con_autovalores
7  )
8  from largevar.simulation import simular_vecm_ec3
9  from largevar.utils import comparar_metodos_estimacion  # si está
       definida
10
11 # Parámetros
```

```
T, N = 500, 10
X, beta_true, alpha_true, _ = simular_vecm_ec3(T, N)

# 1. Calcular autovalores y determinar rango con LargeVAR
eigenvals = compute_eigenvalues_largevar(X)
r_est = test_largevar_rango(X)
print(f"Rango estimado por LargeVAR: {r_est}")

# 2. Estimar beta y alpha con LargeVAR (usando r=1 si el test lo sugiere)
if r_est >= 1:
    beta_lv, alpha_lv = estimar_largevar_con_autovalores(X, eigenvals, r=1)
    print("Beta estimado (LargeVAR):", beta_lv)
    print("Alpha estimado (LargeVAR):", alpha_lv)

# 3. Comparar con Johansen (solo si N es pequeño)
if N <= 20:
    comp = comparar_metodos_estimacion(X, r=1, p=0)
    if 'similitud' in comp:
        print("Correlación entre betas:", comp['similitud']['correlacion'])

# 4. Graficar el spread estimado
spread_lv = X[:, 0] - beta_lv[1] * X[:, 1]  # asumiendo beta normalizado
plt.plot(spread_lv)
plt.title("Spread estimado con LargeVAR")
plt.show()
```

Listing A.6: Ejemplo de uso completo

A.6. Validación y reproducibilidad

Todo el código ha sido probado con las 100,000 simulaciones Monte Carlo para cada valor de N/T en ambos regímenes dimensionales, las mismas que sustentan los resultados del Capítulo 6. Los scripts para generar figuras y tablas están en el directorio `experimentos/`, junto con un archivo `config.yaml` que permite replicar las condiciones exactas de cada simulación. Para reproducir los resultados de la tesis, basta con ejecutar:

```
cd experiments
python run_simulations.py --config config.yaml
```

```
python generate_figures.py
```

Mi intención al publicar todo esto es que cualquier investigador pueda verificar los resultados y, ojalá, extenderlos.

A.7. Licencia

El código se distribuye bajo la licencia MIT. Eso significa que se puede usar, modificar y redistribuir con total libertad, siempre que se mantenga el aviso de copyright original. Si el código llega a utilizarse en investigaciones derivadas, agradeceré que se cite esta tesis como referencia.

Referencias Bibliográficas

[1] R. N. Mantegna y H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge, UK: Cambridge University Press, 1999.

[2] R. F. Engle y C. W. J. Granger, «Co-integration and error correction: Representation, estimation, and testing,» *Econometrica*, vol. 55, n.º 2, págs. 251-276, 1987.

[3] E. Gatev, W. N. Goetzmann y K. G. Rouwenhorst, «Pairs trading: Performance of a relative-value arbitrage rule,» *Rev. Financ. Stud.*, vol. 19, n.º 3, págs. 797-827, 2006.

[4] S. Johansen, «Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models,» *Econometrica*, vol. 59, n.º 6, págs. 1551-1580, 1991.

[5] S. Johansen, *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford, UK: Oxford University Press, 1995.

[6] Z. Bai y J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, 2nd. New York, NY, USA: Springer, 2010.

[7] A. Bykhovskaya y V. Gorin, «Cointegration in large VARs,» *Ann. Stat.*, vol. 50, n.º 3, págs. 1593-1617, 2022.

[8] T. Talasmaa, «Profitability of distance-based pairs trading in Finnish stock markets: 2020-2024,» Tesis de mtría., Aalto University, Espoo, Finland, 2024.

[9] M. Mudelsee, *Climate Time Series Analysis: Classical Statistical and Bootstrap Methods*, 2nd. Cham, Switzerland: Springer, 2019.

[10] G. E. P. Box, G. M. Jenkins, G. C. Reinsel y G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th. Hoboken, NJ, USA: Wiley, 2015.

[11] H. Kantz y T. Schreiber, *Nonlinear Time Series Analysis*, 2nd. Cambridge, UK: Cambridge University Press, 2004.

[12] R. S. Tsay, *Analysis of Financial Time Series*, 2nd. Hoboken, NJ, USA: Wiley, 2005.

[13] P. J. Brockwell y R. A. Davis, *Introduction to Time Series and Forecasting*, 3rd. Cham, Switzerland: Springer, 2016.

[14] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd. New York, NY, USA: Wiley, 2003.

[15] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York, NY, USA: Wiley, 1982.

[16] C. W. J. Granger y P. Newbold, «Spurious regressions in econometrics,» *J. Econom.*, vol. 2, n.º 2, págs. 111-120, 1974.

[17] S. Johansen, «Statistical analysis of cointegration vectors,» *J. Econ. Dyn. Control*, vol. 12, n.º 2-3, págs. 231-254, 1988.

[18] M. S. Ho y B. E. Sorensen, «Finding cointegration rank in high dimensional systems using the Johansen test: An illustration using data based Monte Carlo simulations,» *Rev. Econ. Stat.*, vol. 78, n.º 4, págs. 726-732, 1996.

[19] E. P. Wigner, «Characteristic vectors of bordered matrices with infinite dimensions,» *Ann. Math.*, vol. 62, n.º 3, págs. 548-564, 1955.

[20] L. Laloux, P. Cizeau, J.-P. Bouchaud y M. Potters, «Noise dressing of financial correlation matrices,» *Phys. Rev. Lett.*, vol. 83, n.º 7, págs. 1467-1470, 1999.

[21] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral y H. E. Stanley, «Universal and nonuniversal properties of cross correlations in financial time series,» *Phys. Rev. Lett.*, vol. 83, n.º 7, págs. 1471-1474, 1999.

[22] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr y H. E. Stanley, «Random matrix approach to cross correlations in financial data,» *Phys. Rev. E*, vol. 65, n.º 6, pág. 066 126, 2002.

[23] I. M. Johnstone, «Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy-Widom limits and rates of convergence,» *Ann. Stat.*, vol. 36, n.º 6, págs. 2638-2716, 2008.

[24] A. T. James, «Distributions of matrix variates and latent roots derived from normal samples,» *Ann. Math. Stat.*, vol. 35, n.º 2, págs. 475-501, 1964.

[25] K. W. Wachter, «The limiting empirical measure of multiple discriminant ratios,» *Ann. Stat.*, vol. 8, n.º 5, págs. 937-957, 1980.

[26] A. Soshnikov, «Universality at the edge of the spectrum in Wigner random matrices,» *Commun. Math. Phys.*, vol. 207, n.º 3, págs. 697-733, 1999.

[27] T. Tao y V. Vu, «Random matrices: Universality of local eigenvalue statistics,» *Acta Math.*, vol. 206, n.º 1, págs. 127-204, 2011.

[28] L. Erdős y H.-T. Yau, «Universality of local spectral statistics of random matrices,» *Bull. Am. Math. Soc.*, vol. 49, n.º 3, págs. 377-414, 2012.

[29] F. J. Dyson, «The threefold way. Algebraic structure of symmetry groups and ensembles in quantum mechanics,» *J. Math. Phys.*, vol. 3, n.º 6, págs. 1199-1215, 1962.

[30] I. Dumitriu y A. Edelman, «Matrix models for beta ensembles,» *J. Math. Phys.*, vol. 43, n.º 11, págs. 5830-5847, 2002.

[31] C. A. Tracy y H. Widom, «Level-spacing distributions and the Airy kernel,» *Commun. Math. Phys.*, vol. 159, n.º 1, págs. 151-174, 1994.

[32] C. A. Tracy y H. Widom, «On orthogonal and symplectic matrix ensembles,» *Commun. Math. Phys.*, vol. 177, n.º 3, págs. 727-754, 1996.

[33] I. M. Johnstone, «On the distribution of the largest eigenvalue in principal components analysis,» *Ann. Stat.*, vol. 29, n.º 2, págs. 295-327, 2001.

[34] J. Baik, G. B. Arous y S. Péché, «Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices,» *Ann. Probab.*, vol. 33, n.º 5, págs. 1643-1697, 2005.

[35] X. Han, G. Pan y B. Zhang, «The Tracy-Widom law for the largest eigenvalue of F type matrices,» *Ann. Stat.*, vol. 44, n.º 4, págs. 1564-1592, 2016.

[36] M. Chiani, «Distribution of the largest eigenvalue for real Wishart and Gaussian random matrices and a simple approximation for the Tracy-Widom distribution,» *J. Multivar. Anal.*, vol. 129, págs. 69-81, 2012.

[37] A. Bejan, «Largest eigenvalues and sample covariance matrices. Tracy-Widom and Painlevé II: Computational aspects and realization in S-Plus with applications,» Preprint, University of Cambridge, 2005.

[38] A. Onatski y C. Wang, «Alternative asymptotics for cointegration tests in large VARs,» *Econometrica*, vol. 86, n.º 4, págs. 1465-1478, 2018.

[39] A. Onatski y C. Wang, «Extreme canonical correlations and high-dimensional cointegration analysis,» *J. Econom.*, vol. 212, n.º 1, págs. 307-322, 2019.

[40] G. Vidyamurthy, *Pairs Trading: Quantitative Methods and Analysis*. Hoboken, NJ, USA: Wiley, 2004.

[41] D. H. Bailey y M. L. de Prado, «The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality,» *J. Portf. Manag.*, vol. 40, n.º 5, págs. 94-107, 2014.