LA IMPUTACIÓN MÚLTIPLE Y SU APLICACIÓN A SERIES TEMPORALES FINANCIERAS

Sebastian Cano Berlanga

Dipòsit Legal: T.76-2014

UNIVERSITAT ROVIRA I VIRGILI

TESIS DOCTORAL

La imputación múltiple y su aplicación a series temporales financieras

PRESENTADA POR SEBASTIÁN CANO BERLANGA

Tutora: Dra. María José Pérez Lacasta

Director: Dr. Máximo Borrell Vidal

Departamento de Economía
Facultad de Economía y Empresa
Reus, Octubre 2013

UNIVERSITAT ROVIRA I VIRGILI
LA IMPUTACIÓN MÚLTIPLE Y SU APLICACIÓN A SERIES TEMPORALES FINANCIERAS
Sebastian Cano Berlanga
Dipòsit Legal: T.76-2014

"Hay ocasiones en las que,
para que la cercanía florezca,
es necesario cultivar la distancia."

UNIVERSITAT ROVIRA I VIRGILI
LA IMPUTACIÓN MÚLTIPLE Y SU APLICACIÓN A SERIES TEMPORALES FINANCIERAS
Sebastian Cano Berlanga
Dipòsit Legal: T.76-2014

Agradecimientos

Esta tesis doctoral refleja la culminación del estudio llevado a cabo durante los últimos años, sin embargo la consecución de este objetivo no se debe únicamente a la investigación, sino a las maravillosas relaciones establecidas con personas extraordinarias.

En primer lugar me gustaría agradecer a la Facultad de Economía y Empresa y al departamento de Economía el apoyo mostrado a lo largo de estos años. En segundo lugar me gustaría mostrar mi gratitud a:

Dr. Bernd Theilen, director del departamento de Economía, a la Dra. Carolina Manzano por sus ánimos y motivación, y al Dr. Jordi Andreu por su compañía incondicional en los buenos y malos momentos.

Mi más sincero agradecimiento a la Dra. María José Pérez, tutora de esta Tesis, y a mi director, el Dr. Máximo Borrell, que no sólo ha dirigido, ha sido un compañero de viaje sin el cual el presente trabajo no podría haberse culminado.

Por último doy las gracias a mis padres, pilares fundamentales para que esta Tesis siguiera adelante.

Índice general

Índice de cuadros

Índice de figuras

UNIVERSITAT ROVIRA I VIRGILI
LA IMPUTACIÓN MÚLTIPLE Y SU APLICACIÓN A SERIES TEMPORALES FINANCIERAS
Sebastian Cano Berlanga
Dipòsit Legal: T.76-2014

Introducción

"Más vale encender una vela, que maldecir la oscuridad"
Proverbio popular chino

Esta *Introducción* tiene por objeto definir y acotar el tema de investigación así como describir la organización del trabajo que se presenta. Precedemos la definición y acotación del tema por una reflexión etimológica primero, y, seguidamente, por la exposición del recorrido bibliográfico que nos ha conducido a esa finalidad.

Consideraciones etimológicas

En la investigación empírica es muy común encontrar bases de datos incompletas, es decir, aquéllas cuyos valores *no disponibles*[1] (NA, *not available*) condicionan el uso de las técnicas convencionales de análisis estadístico.[2]

La preocupación por el problema de los NA ha dado lugar a una amplia literatura debido a que para la realización de análisis estadísticos la presencia de NA (descartada su omisión) conlleva la aparición de tres obstáculos:[3] (i) imposibilidad de ser tratados

[1] En castellano no hay un consenso para la traducción del término *missing*, utilizándose indistintamente "valor faltante", "valor perdido", "dato ausente", etc. Creemos que la manera más correcta de traducir *missing* es "valor no disponible", representado por las siglas NA (nomenclatura que, además, es la que se utiliza en **R**); y la creemos más correcta porque un valor NA puede que, efectivamente, se haya "perdido", pero también que"no esté disponible" debido a otras causas ("no obtenido todavía", "sin posibilidad de existencia", etc) . En esta Tesis usamos las siglas NA para referirnos, por tanto, a las celdas vacías de una base de datos.

[2] En la construcción tradicional de la Estadística se parte implícitamente de la completitud de la base de datos; sin embargo, en la actualidad, el manejo y construcción de bases de datos es tan generalizado que, incluso, ha dado lugar al nacimiento de nuevas ramas de la Estadística, como el *data mining*. Esto ha inducido la necesidad de utilizar técnicas que subsanen los problemas derivados de la falta de datos y, así, estar en condiciones de proseguir el análisis de una manera eficiente.

[3] Según Groves [1989], la presencia de valores NA puede ser causada por: (i) errores de cobertura y (ii) falta de respuestas, lo que supone la imposibilidad de obtener información total o parcial de un determinado registro. Esta definición tan genérica sobre la falta de respuesta está ampliada en Vach [1994], quien ilustra con una variedad de ejemplos las posibles causas de la falta de respuesta.

informáticamente mientras no se les asigne un valor; (ii) elección del procedimiento algorítmico de imputación; y (iii) sesgo introducido por el algoritmo elegido.

Consultemos qué dicen los diccionarios acerca de las voces *imputación* y *plausibilidad*. Según el Diccionario de la Real Academia Española: *Imputación:* [1] Acción y efecto de imputar. *Imputar:* en su segunda acepción "Señalar la aplicación o inversión de una cantidad, sea al entregarla, sea al tomar razón de ella en cuenta". A nuestro juicio matiza mejor la definición el diccionario de María Moliner, **que también en su segunda acepción** nos dice: "Asignar cierto destino a una cantidad, al entregarla o al consignarla". De esta definición hemos de destacar el verbo *asignar* y la expresión *cierto destino a una cantidad*. En el sentido que manejamos, "destino" sería la prosecución del análisis estadístico, y "asignación" sería el resultado de la imputación.

En su definición estadística más amplia, la imputación es la asignación de uno o más valores **plausibles**, que permitan efectuar un eficiente análisis de datos. La figura 1 describe lo que a nuestro juicio son las posibilidades existentes a la hora de enfrentarse a una base de datos incompleta.

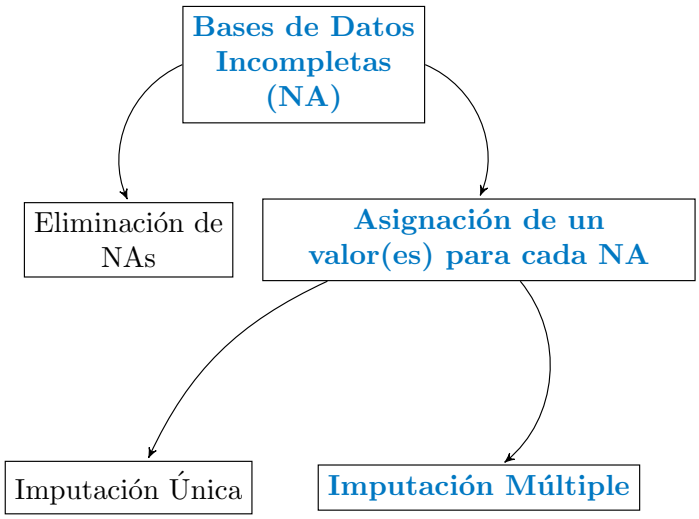

FIGURA 1: Alternativas para enfrentarse a bases de datos incompletas. La eliminación de NA es factible si éstos se disponen aleatoriamente y, además, su número es inferior al 5 %; en caso contrario, han de utilizarse técnicas de imputación, sean de imputación única o múltiple, Schafer [1997].

La clave del éxito de cualquier técnica de imputación es, pues, la plausibilidad del valor asignado ya que, de lo contrario, el valor dado a los NA debilitará el ulterior análisis de la base de datos. Debido a la importancia que, en el ámbito de la imputación, posee

el término *plausible*, consideramos conveniente exponer una reflexión semántica sobre el mismo:

Real Academia Española: [1] Digno o merecedor de aplauso. [2] Atendible, admisible, recomendable.

María Moliner: [1] (aplicado a acciones) Digno de aplauso o alabanza. [2] (aplicado a motivos o razones) justificado, admisible o atendible.

Ambas definiciones coinciden en que *plausible* significa que tiene mérito, por tanto digno de alabanza, y que además se trata de algo razonable, por consiguiente, admisible o recomendable. La palabra *plausible* admite mayores matices en lengua inglesa, como comprobamos a continuación (traducción propia):

Diccionario Webster: [1] En apariencia, justo, razonable o valioso. [2] Superficialmente placentero o persuasivo. [3] Aparentemente merecedor de ser creído.

Diccionario de Oxford: [1] (referido a un argumento o afirmación) Con apariencia de razonable o probable. [2] (referido a una persona) Que es capaz de elaborar argumentos persuasivos, especialmente aquéllos con intención de engañar.

Diccionario de Cambridge: [1] Con apariencia de verdadero, o capaz de ser creído. [2] (en sentido negativo) Describe a alguien aparentemente honrado y sincero.

Estos diccionarios están en línea con las definiciones dadas en nuestros diccionarios y, además, añaden los matices de "persuasivo" y de "aparentemente". Los diccionarios ingleses introducen la posibilidad de usar la plausibilidad como herramienta para engañar, confundir y manipular.[4] Es decir, la *acción* de imputación conlleva la *apariencia* de verdadero o creíble, lo que sugiere una eventual contrastación. Por consiguiente, en nuestro trabajo, cuando hablamos de imputación estadística nos referimos a:

> una **acción** (imputadora), bien motivada o razonada, para la cual se establece un criterio (de imputación) y se elige una técnica que posibilite calificar los resultados como plausibles (contrastación).

En el proceso de definición y acotación del tema de investigación arrancamos de este concepto.

En tanto que materia prima para cualquier imputación estadística, los datos deberían reunir un conjunto de características ideales. Desde una perspectiva genérica, Redman [1992] indica que, para ser considerados ideales, los datos deben tener las siguientes características:

[4] Dempster & Rubin [1983] advierten que las técnicas de imputación pueden ser tan seductoras como peligrosas y aun expresamente engañosas ya que una mala utilización distorsiona los resultados del análisis.

Precisión, o medida de la diferencia entre el valor verdadero y el valor registrado.

Plenitud, o proporción de NA con respecto al total de los datos observados en el dataset, estimada conveniente.

Actualidad, o influencia del paso del tiempo en los datos.

Consistencia, o coherencia interna de la base de datos.

En el caso de que los datos analizados no reúnan estas características, puede que se realice un trabajo analítico poco eficiente, situación para la que el idioma inglés tiene un acrónimo, GIGO (*garbage in, garbage out*). La presente Tesis aborda la *plenitud* de los datos, e intenta que la solución propuesta cumpla con las tres características restantes.

Definición y acotación del tema de investigación

Realicemos ante todo unas consideraciones acerca del análisis de datos. El análisis estadístico de datos completos puede permitir la correcta estimación de los parámetros de un modelo lineal $y = \beta \boldsymbol{X} + \varepsilon$, en cuyo caso, si el término de perturbación es gaussiano (media nula y varianza constante), la mejor estimación de β la proporciona la siguiente ecuación, que se utiliza cuando la matriz de datos ($\boldsymbol{X}$) está completa:[5]

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t y$$

en el caso de que la matriz de datos contenga NA, es de esperar que el resultado de la estimación del vector β cambie, pues las matrices de datos son distintas.

La solución inmediata consistiría en eliminar los NA de la base de datos; de esta forma, se obtendría un registro completo, aunque de menor tamaño muestral, siendo posible entonces aplicar los métodos típicos de análisis. Sin embargo, esta forma de proceder causa una reducción del rango de la matriz de datos, con el consiguiente descenso del número de grados de libertad, así como una inferencia estadística menos precisa, dando lugar a una peor estimación de los parámetros poblacionales.

En caso de que se desee restablecer el rango original de la matriz de datos (obviamente, bajo la hipótesis de que la información no disponible puede ser completada), cabe admitir que sea posible elaborar una técnica que permita completar la base de datos $\boldsymbol{X}$ (incompleta), de manera que se obtenga una matriz X^* (**imputada**), y, que a su vez sea coherente con la base de datos $\boldsymbol{X}$ (completa). De ahí arranca el concepto de técnica de **imputación** o, simplemente imputación, como aquélla que permite asignar un valor plausible a un NA; una vez realizada la totalidad de las asignaciones podemos visualizar así la relación entre $\boldsymbol{X}, X^*$ y $\boldsymbol{X}$.

[5] Ver, por ejemplo, Gujarati [1997, p. 283].

$$\boldsymbol{X} \xrightarrow{imputación} X^* \xrightarrow{coherente\ con} \boldsymbol{X}$$

Nuestra Tesis doctoral pretende aplicar la imputación múltiple a series temporales financieras univariantes de carácter bursátil (STF), para ello nos guiamos por la estructura secuencial que acabamos de decribir, con la salvedad de los pertinentes cambios notacionales.Obviamente, tuvimos que efectuar un recorrido bibliográfico que condujera a la determinación del objeto, objetivos y límites de la investigación.

Aceptado el concepto de plausibilidad que hemos dado, se presentó el problema de cómo realizar una imputación múltiple sobre un dataset formado por STF y cómo contrastar los resultados proporcionados por la misma. Sintéticamente, el citado recorrido bibliográfico se expone en el diagrama siguiente (figura 2):

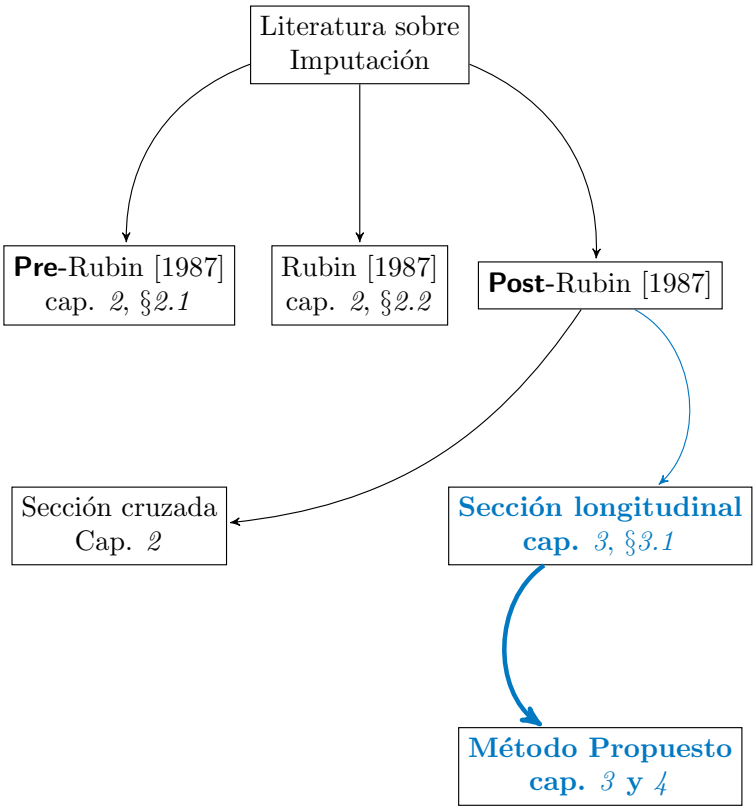

FIGURA 2: Recorrido bibliográfico sobre imputación

* * *

Hipótesis requeridas sobre los datos de una serie temporal para realizar su análisis:

1. Intervalos regulares de tiempo.
2. No existen valores NA.
3. Número mínimo de valores ($\geqslant 50$)
4. La serie temporal es estacionaria, es decir:
   - sin tendencia
   - sin estacionalidad
   - el carácter del ruido (amplitud y frecuencia) no cambia durante el periodo observado.

En muchas series temporales reales fallan una o más de estas hipótesis, lo que obliga a realizar una limpieza de datos antes de proceder a su análisis. La limpieza de datos significa que:

- si los datos han sido muestreados a intervalos irregulares de tiempo, quizá se pueda intentar un remuestreo a intervalos constantes;
- si el número de valores es < 50 (cosa que ocurre a menudo en la práctica empresarial, en la que se suele disponer, de a lo sumo dos docenas de datos), usar la suavización exponencial, por tratarse de un filtro para reducir la magnitud del ruido;
- si la serie temporal no es estacionaria, se debería identificar y eliminar las componentes de tendencia (a largo plazo) y estacionalidad;
- si se producen cambios profundos en la naturaleza de la serie temporal (debido, por ejemplo, a un suceso "rompedor") lo procedente es escindir la serie temporal en otras dos, una **antes** del suceso y otra **después** del mismo;
- si existen valores NA en número suficientemente elevado, en lugar de prescindir de ellos, se plantea la necesidad de utilizar una técnica que permita asignar valores.

En lo que se refiere a esta asignación, el recorrido bibliográfico efectuado mostró que la imputación múltiple ha sido aplicada únicamente sobre datos de sección cruzada. *Esto llevó a interrogarnos acerca de la existencia de una literatura sobre valores NA en STF utilizando la técnica de imputación múltiple*, ya que cabe que haya datos no registrados debido a varias causas, por ejemplo, desastres naturales, guerras, ausencia de financiación o suspensión de cotización, entre otras.

Modelo	**Artículos**
ARMA	Ferreiro [1987] y Parzen [1984]
Espacio de estados	Durbin & Koopman [2004]
Suavización	Jong [1995]; He, Yucel & Raghunathan [2011]
Filtro de Kalman	Kalman [1960]; Harvey & Pierse [1984]
Redes Bayesianas	Pearl [1988, 2009]
Redes neuronales	Alexiadis et al. [1998]; Kihoro et al. [2007]
Patrones estacionales	Chiewchanwattana, Lursinsap & Chu [2007]
Algoritmos genéticos	Figueroa García, Kalenatic & Lopez Bello [2008]

Cuadro 1: Modelos utilizados para imputar series temporales mencionados en Denk & Weber [2011]

* * *

En principio, los métodos de imputación aplicables a sección cruzada lo son también en el contexto de series temporales tal como vemos en Denk & Weber [2011], quienes realizan un recorrido bibliográfico sobre los métodos de imputación en series temporales, enfatizando, sin embargo, *que el éxito de las distintas técnicas de imputación depende del patrón de valores NA que presente la serie temporal.*[6]

El trabajo de Denk & Weber [2011] pone de manifiesto la existencia de una gran variedad de alternativas para imputar series temporales (para un listado de los métodos mencionados en dicho trabajo, ver cuadro 1). Para estos autores, la forma más sencilla de enfrentarse al problema consiste en "llevar el último valor hacia adelante", es decir, utilizar la última observación conocida para completar la serie temporal; pero indican también que puede utilizarse la media aritmética para las asignación de valores NA. Ambos métodos *introducen un sesgo, pues alteran la distribución centrándola en un determinado valor.*[7]

* * *

[6]Según esos autores los patrones más relevantes relativos a series temporales univariantes son: (i) *NA esporádicos*, donde los valores NA se disponen aleatoriamente, (ii) *NA de periodo completo*, donde los *NA* se hallan durante periodos temporales determinados.

[7]Los mismos autores sugieren que pueden utilizarse métodos de interpolación sencillos, como medias móviles, o más sofisticados, como los splines. Como inidcan Ferreiro [1987] y Parzen [1984], la elección más frecuente para completar series temporales consiste en utilizar modelos AR, MA o ARMA. También es posible aplicar modelos de espacio de estados, Durbin & Koopman (2004) o algoritmos de suavización, Jong [1995] y He, Yucel & Raghunathan [2011]. Resulta digno de consideración recordar que Denk & Weber (2011] destacan que la imputación de series temporales no se restringe a los modelos estadísticos más usuales, sino que además se encuentran técnicas basadas en: redes neuronales, Alexiadis et al. [1998]; Kihoro et al. (2007); identificación de patrones estacionales, Chiewchanwattana, Lursinsap & Chu (2007); y algoritmos genéticos y heurísticos, Figueroa García, Kalenatic & Lopez Bello (2008).

La revisión de la literatura nos condujo a las siguientes conclusiones:

- existe literatura sobre valores NA con imputación múltiple en datos sección cruzada;
- existe literatura sobre valores NA en series temporales;
- **no hemos encontrado literatura sobre imputación múltiple aplicada a series temporales financieras univariantes**.

Debido a la inexistencia de la literatura específica, nuestra investigación ha tenido que seguir el camino emprendido por Andreu & Cano [2008] y Cano & Andreu [2010], lo que significa que hemos debido utilizar nuestras propias fuerzas para tratar el tema, y ello tanto desde el punto de vista teórico como en el de programación informática (ambas en el capítulo 3). Estos dos artículos constituyen una base preliminar para la presente investigación, ya que sin ellos no podría haber sido posible la profundización necesaria para aplicar la imputación múltiple a series temporales financieras univariantes.

Objetivos de la Tesis Doctoral

En esta Tesis Doctoral nos formulamos los siguientes objetivos:

- Objetivo 1: En la literatura existe un *gap* entre la imputación múltiple aplicada a datasets de sección cruzada y la teoría de las STF. Esto determina el objetivo principal del trabajo: *diseñar un método que enlace la teoría de la imputación múltiple y el ámbito de las series temporales financieras bursátiles, considerando el tiempo como única variable independiente.*
- Objetivo 2: Generar *conjuntos de valores plausibles* para STF mediante el uso de los algoritmos *Gibbs Sampling* y *Approximate Bayesian Bootstrap* (ABB).
- Objetivo 3: Partiendo del conocimiento de la inferencia de Rubin, adaptarla informáticamente para poder construir intervalos de confianza sobre los coeficientes de los modelos usuales de series temporales financieras.
- Objetivo 4: Aplicar empíricamente el método propuesto y contrastar la *plausabilidad* de los valores simulados. Dicha contrastación se llevará a cabo sobre el análisis comparado de los coeficientes estimados de tres modelos GARCH. En este entorno entendemos que la plausibilidad es satisfactoria si se cumplen las tres siguientes condiciones:

1. el contraste de significación de los coeficientes calculados mediante imputación múltiple coincide con el contraste de significación cuando el dataset no contiene valores NA;
2. el signo de los coeficientes calculados mediante imputación múltiple coincide con el signo cuando el dataset no contiene valores NA
3. las magnitudes de ambos coeficientes son similares;

- Objetivo 5: Implementar el método propuesto en lenguaje **R**. Este objetivo tiene 3 fases:

  1. Instrucciones que introducen valores NA en y_t
  2. Instrucciones que generan valores plausibles (etapa de *imputación*)
  3. Instrucciones que implementan las etapas de *análisis* y *combinación*

  Las instrucciones que hemos programado al efecto constituyen la librería **mists** (*Multiple Imputation Simulation for Time Series*)

Metodología utilizada

El tratamiento de una STF, y_t, como un todo impide que la imputación múltiple tenga éxito, pues la no estacionariedad en varianza altera los resultados producidos por el Gibbs Sampling. Además, la naturaleza de y_t, exige que los valores simulados satisfagan las dos siguientes condiciones:

- carácter asimétrico de la volatilidad (Black [1976], Taylor [1986])
- presencia de colas pesadas (Mandelbrot [1963])

En lugar de de tratar una STF como un todo, la búsqueda de la satisfacción de ambas condiciones nos conduce a separar y_t mediante un filtro GARCH *asimétrico*, que la descompone multiplicativamente en dos procesos:

- **volatilidad** (σ_t), para capturar la asimetría
- **innovaciones** (ε_t), para capturar la leptocurtosis

Esta es, precisamente, la idea clave que proponemos: *escindir la serie original en dos partes, para posteriormente generar una serie temporal imputada y plausible.* Intentando desarrollar esta idea estudiamos distintas herramientas analíticas ya presentes en la

literatura. Felizmente, encontramos que el Threshold GARCH, y los algoritmos Gibbs Sampling y el ABB[8] podían ser utilizados al servicio de la estrategia de generación de valores imputados plausibles. La citada idea clave, como arco de bóveda, en conjunción con las herramientas analíticas indicadas configura lo que aquí denominamos como un nuevo procedimiento que denominaremos método de **imputación mediante separación**.

La separación multiplicativa de y_t en σ_t y ε_t, permite tratar estas componentes diferentemente: *Gibbs Sampling* para imputar las innovaciones, y *ABB* para imputar la volatilidad.

La aplicación concreta del nuevo procedimiento a un conjunto de STF previamente elegidas la realizaremos en el capítulo 4, donde nos valdremos de un artificio: tomaremos un conjunto de STF completas y generaremos diversos escenarios en forma de porcentajes de NA, trabajando a continuación sobre dichos escenarios.

Medios utilizados para alcanzar los objetivos

Como base para la consecución de los cinco objetivos anteriormente indicados se han utilizado las siguientes herramientas:

- Instrumental analítico específico
  - Cadenas de Markov
  - Simulación de Montecarlo
  - *Gibbs Sampling* (un algoritmo MCMC)
  - *Approximate Bayesian Bootstrap* (un método Bootstrap)
  - *Threshold GARCH*
- Instrumental informático

  R: Lenguaje y entorno de programación libre para análisis estadístico y gráfico con librerías específicas para el análisis de datos ante la presencia de NA.

[8] Podíamos haber utilizado otros algoritmos alternativos, como el *Data Augmentation* de Tanner & Wong [1987], o el *Bootstraped EM* de Honeker et al. [2011]. Decidimos utilizar el *Gibbs Sampling* por su uso generalizado en la literatura de imputación múltiple, y el *ABB* debido a su sencillez y eficacia computacional.

Organización del trabajo que se presenta

La Tesis contiene la presente *Introducción*, cuatro capítulos, *Conclusiones* y *Bibliografía*.

Los dos primeros capítulos tienen por objeto describir los instrumentos necesarios para pasar a los capítulos *3* y *4*.

Capítulo 1: Expone el lenguaje **R**, así como un resumen del instrumental analítico sobre el cual descansa la técnica de imputación múltiple: cadenas de Markov, simulación de Montecarlo y Gibbs Sampling.

Capítulo 2: Expone la aplicación de la imputación múltiple en el ámbito de la sección cruzada como herramienta para asignar conjuntos de valores plausibles a los NA. Tras el correspondiente recorrido bibliográfico se expone de forma general la técnica de imputación múltiple, centrándonos en la inferencia de Rubin [1987], así como las limitaciones de ésta. Cerramos el capítulo con un ejemplo que pone de manifiesto el funcionamiento de la técnica.

Los dos capítulos siguientes tienen por finalidad describir, implementar y contrastar el modelo de imputación mediante separación que proponemos.

Capítulo 3: Se describe el método que presentamos para atacar el problema de la imputación múltiple en series temporales financieras univariantes. El desarrollo del método que proponemos requiere el uso auxiliar de dos instrumentos: Threshold GARCH y ABB. La implementación informática del método propuesto nos llevó a la construcción de la librería específica que denominamos **`mists`**, programada en **R**. Existen otras librerías que permiten utilizar imputación múltiple en **R**, pero ninguna de ellas funciona con datos de sección longitudinal, como son las STF que tratamos.

Capítulo 4 Aplicación empírica y evaluación del método propuesto sobre una muestra de firmas cotizadas y de índices de mercado que elegimos al efecto.

El capítulo de **Conclusiones** se desarrolla según una doble óptica:

- *Conclusiones Retrospectivas:* aquéllas que se derivan directamente del propio trabajo.
- *Conclusiones Prospectivas:* líneas futuras de investigación que a nuestro juicio se desprenden de la Tesis.

Finalmente, se incluye la **Bibliografía**, que se refiere a la citada y a la vez consultada, exceptuando los diccionarios españoles de uso corriente (RAE y María Moliner) y los ingleses (Webster, Oxford y Cambridge).

Capítulo 1

Instrumentos

Este capítulo tiene por finalidad exponer los instrumentos de base (informático y analíticos) que utilizamos para la formulación del modelo que proponemos en los capítulos *3* y *4*.

Como instrumento informático elegimos el lenguaje **R**, orientado a objetos, para el cual daremos una descripción general, y expondremos los rasgos esenciales de la programación de *instrucciones*; culminaremos la presentación de **R** con un código que calcula de forma sencilla las ponderaciones de activos bursátiles según el conocido modelo de Markowitz, tomando como input una matriz de precios previamente descargada de internet a través del propio entorno.

En lo que se refiere a los instrumentos analíticos se exponen sucintamente aquéllas herramientas que se utilizan para implementar la técnica de imputación múltiple: *cadenas de Markov*, *simulación de Montecarlo* y *MCMC*.

1.1. El lenguaje R

R es un lenguaje y entorno de programación para análisis estadístico[1] creado por Robert Gentleman y Ross Ihaka, del Departamento de Estadística de la Universidad de Auckland (Nueva Zelanda), que se sirvieron de la inicial de sus nombres de pila para dar nombre al programa.[2] Se trata de un proyecto de software libre, resultado de la implementación GNU del lenguaje **S**, puesto a disposición del público en su versión 1.0.0 en al año 2000. Los paquetes R y S-Plus son, probablemente, dos de los lenguajes más utilizados por la comunidad estadística, por ejemplo en los campos de la investigación Biomédica, Bioinformática y de la matemática de los mercados financieros.

[1] Un resumen de las características principales de **R** pueden verse en el cuadro 1.1.

[2] Existen numerosos programas para realizar análisis estadísticos y matemáticos: Stata, SAS, SPSS, eViews, Matlab y Mathematica entre muchos otros, pero nos hemos decantado por **R** debido a que sus ventajas resultaron las más atractivas para nuestro trabajo.

Característica	Comentario
Libre	Gratuito, de código abierto (modificable directamente por el propio usuario) y puede instalarse en cualquier ordenador gracias a su diseño de plataforma cruzada.
Documentación	Dispone de una gran cantidad recursos de documentación de carácter académico y profesional en constante crecimiento.
Rápida expansión	Su comunidad de usuarios es muy activa, y difunde rápidamente las novedades mediante la etiqueta de twitter **#rstats** y la página *www.r-bloggers.com*.
Elevado número de librerías	Numerosas ramas científicas tienen librerías específicas para aplicar sus modelos. Están especialmente desarrolladas la Estadística Bayesiana, la Bioestadística y Finanzas.
Funciona con línea de comandos	Se mejora así la comprensión de las instrucciones, es más rápido repetir su ejecución y facilita el almacenamiento.
Potencia gráfica	Dispone de un potente motor de gráficos. Un ejemplo de esta potencia son las librerías **ggplot2**, **lattice** y **googleVis**.
Importación y exportación	R es capaz de leer y escribir en todos los formatos disponibles de datos, tanto libres como comerciales (Stata, SPSS, SAS, Matlab,Excel,...) gracias a la librería **foreign**.
Interactúa con LaTeX	R Exporta tablas de resultados a LaTeX mediante la librería **xtable** y escribir el código de los gráficos utilizando **tikzDevice**.
Utiliza directamente URLs	Puede trabajar directamente datos ubicados en internet mediante la especificación de su URL.
Conexión con servidores de Finanzas	Descarga los datos financieros disponibles en los portales de Yahoo, Google, y Oanda.

Cuadro 1.1: Características principales de **R**

R trae de serie en el paquete *base* una amplia gama de modelos estadísticos y capacidades gráficas, que incluye modelos lineales y no lineales, contrastes de hipótesis, series temporales y análisis cualitativo entre otros; además, posee unas posibilidades de importación y exportación muy interesantes, como puede verse a continuación,

Programas estadísticos	Gretl, SAS, SPSS, Stata, Excel
Ficheros de texto	ASCII, XML, HTML, CSV
Bases de datos	Access, MySQL, SQL, Oracle

Posibilidades de importación en R

El lenguaje **R** es fácilmente extensible mediante la adición de librerías contribuidas, por lo que es posible utilizar el programa en numerosos ámbitos. La instalación de paquetes contribuidos se consigue mediante a la utilización del comando `install.packages()`, que permite descargar las librerías especificadas y sus *dependencias* e incorporarlas en el sistema; incluyendo además la documentación referente a su uso.

Gracias a su diseño multiplataforma, podemos usar **R** en cualquier sistema operativo, lo que ofrece la ventaja de que los usuarios de **R** puedan intercambiar materiales y programas de forma automática sin tener que adaptar el código a las peculiaridades del sistema operativo. Los resultados obtenidos en **R** pueden exportarse a LaTeX, y los gráficos en formatos vectoriales (por ejemplo PDF) lo que ahorra tiempo en la edición de documentos, proporcionando resultados profesionales.

Otro punto fuerte del programa es su extensa documentación, accesible mediante el comando `help.start()`, que ofrece una completa descripción de las características básicas de **R**, documentación específica de todas las librerías instaladas, y funciones de búsqueda. También es de destacar el impulso que ha recibido **R** en numerosísimas publicaciones, desde su inicio hasta la actual versión 3.

La comunidad de usuarios de **R** es muy activa y dinámica, y tiene su propio lugar en internet. Las novedades se difunden rápidamente en twitter mediante la etiqueta **#rstats**, y las entradas más relevantes de los blogs que hacen referencia a **R** se incluyen en la página **r-bloggers**. También hay conferencias académicas y profesionales centradas en R, como "UseR!" que se reúne bienalmente, o "R/Finance", que congrega anualmente a los usuarios de R dedicados a Finanzas Aplicadas. Existe asimismo la revista de publicación semestral "The R Journal" donde se publican los artículos más relevantes relativos a **R**.

La programación en R

R además de proporcionar las instrucciones comúnmente utilizadas, permite su extensión mediante la programación de instrucciones propias,[3] las cuales se rigen por la siguiente estructura básica:

- **Definición:** asignación de nombre y argumentos de la función, que se realiza mediante `function()`
- **Librerías adicionales:** carga de las dependencias de la instrucción, que se realiza mediante `require()`
- **Contenido:** código principal, que puede incluir condicionales, bucles, gráficos, etc.
- **Objetos intermedios:** almacenamiento de los objetos qué desean guardarse, que se realiza mediante `list()`
- **Salida:** indica a **R** que muestre el resultado de la instrucción, se realiza mediante `return()`

según el esquema:

I:	**nombre** ← *function(argumentos)*{
II:	*require(librerías adicionales)*
III:	CONTENIDO
IV:	**resultado** ← *list(objetos que desean guardarse)*
V:	**return**(*resultado*)}

Para ilustrar la potencia y facilidad de programación en **R** nos permitimos exponer una instrucción que calcula los pesos de los activos de una cartera bursátil de acuerdo con la conocida teoría de Markowitz, y los representa gráficamente.

El comando diseñado transformará las series temporales de precios de los activos en series temporales de rendimientos logarítmicos y luego minimizará el riesgo. La instrucción la hemos llamado `markowitz.portfolio()` y su código es el siguiente(posteriormente se dibujará la frontera eficiente utilizando simulación de Montecarlo):

```
markowitz.portfolio ← function(x){
# variable x: matriz de precios de los activos.
```

[3] La programación de instrucciones en **R** comparte muchas similitudes con la del resto de lenguajes de programación usuales.

```
# Se cargan las librerias adicionales.
require(fPortfolio)
require(timeSeries)
```

```
# Se inicia el contenido principal de la instruscción
diff(log(as.matrix(x))) → x.ret
timeSeries(x.ret) → x.ret
Spec = portfolioSpec(); setTargetReturn(Spec) = NULL;
Constraints = "LongOnly"
cartera ← minvariancePortfolio(x.ret, Spec, Constraints)
```

```
# Se dibuja el gráfico
weightsPie(cartera)
```

```
#guardamos objetos intermedios
resultado ← list(cartera=cartera)
```

```
#salida y cierre de la instrucción
return(resultado)
}
```

Para ejecutar la instrucción `markowitz.portfolio()` vamos a utilizar un dataset que contiene la cotización diaria de cinco acciones: Amazon, Apple, IBM, Intel y Microsoft entre 2010 y 2012. El objeto que contiene esta información es una matriz a la que hemos llamado **activos**. El resultado tras aplicar `markowitz.portfolio()` al objeto **activos** puede verse en la figura 1.1.

```
>markowitz.portfolio(activos)
```

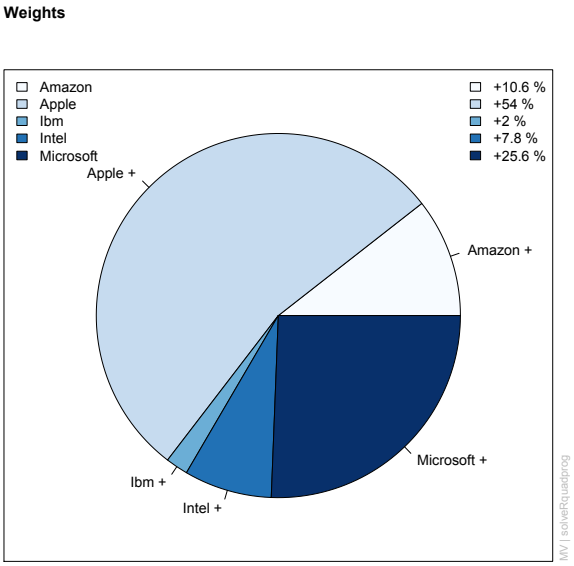

FIGURA 1.1: Representación gráfica de la cartera óptima.

1.2. Instrumentos analíticos

La razón de ser de los instrumentos que aquí tratamos es la siguiente: en primer lugar cadenas de Markov en tiempo discreto y simulación de Montecarlo, necesarias ambas para acceder al conjunto de métodos MCMC; y éste, a su vez, como trampolín para el estudio de la técnica de imputación múltiple.

1.2.1. Cadenas de Markov

El primer elemento necesario para comprender el funcionamiento de los métodos MCMC son las cadenas de Markov, nombrada así en homenaje al gran matemático ruso.[4] La cadena de Markov que aquí nos interesa es un proceso estocástico en tiempo discreto que cumple la importante *propiedad de Markov* (según la cual los estados futuros de una variable aleatoria son independientes de los pasados), cuyo uso se ha ido difundiendo y aplicado a distintos campos: Música, Economía y Finanzas,...,etc;[5] con lo que se pone

[4] Andrey Markov fue un distinguido matemático ruso en el siglo XIX, discípulo de Chebyshev en la Academia de Ciencias de San Petersburgo. Sus contribuciones al campo de la Matemática son importantísimas, en concreto son notables sus investigaciones sobre las fracciones continuas y los procesos estocásticos.

[5] El trabajo original de Markov, aplicado al estudio estadístico del poema *Eugene Oneguin*, puede consultarse en Markov [2006]

de relieve la versatilidad que posee este proceso estocástico.[6]

Definición

Una cadena de Markov es una secuencia de variables aleatorias que satisfacen la propiedad de Markov.[7] Formalmente,

$$Pr(X_{n+1} = X_n | x_n, ..., X_1 = x_1) = Pr(X_{n+1} | X_n = x_n)$$

El proceso de Markov está formado por dos elementos: el vector $\boldsymbol{u}$, en el que se incluyen los estados de la cadena en un determinado momento; y la la matriz $\boldsymbol{P}$, que contiene por las *probabilidades de transición*,[8] que son las probabilidades condicionadas asociadas a instantes temporales sucesivos. Por ejemplo, si la cadena únicamente tiene dos estados, entonces,

$$\boldsymbol{P} = \begin{bmatrix} p_{ii} & p_{ij} \\ p_{ji} & p_{jj} \end{bmatrix}$$

Las matrices de transición verifican los dos siguientes teoremas:

Teorema 1: si $\boldsymbol{P}$ es la matriz de transición de una cadena de Markov, entonces $\boldsymbol{P}^n$ es la probabilidad de que la cadena esté en el estado s_j, partiendo de s_i, tras n pasos.

Teorema 2: si $\boldsymbol{P}$ es la matriz de transición de una cadena de Markov y $\boldsymbol{u}^{(0)}$ es el vector de probabilidad que representa la distribución de estados inicial, entonces $\boldsymbol{u}^{(n)}$ representa el vector $\boldsymbol{u}$ tras n pasos,

$$\boldsymbol{u}^{(n)} = \boldsymbol{u}^{(0)} \cdot \boldsymbol{P}^n$$

Estas propiedades son fundamentales para determinar el estado estacionario de la cadena, que a su vez influye en los algoritmos MCMC.

En las cadenas de Markov cabe la posibilidad de que un estado sea absorbente, y esto sucede si, una vez alcanzado éste, no puede abandonarse a pesar de los cambios que $\boldsymbol{P}$ experimente.[9]. Para que una cadena de Markov sea absorbente se ha de dar una de estas dos condiciones:

[6] Para una introducción a las cadenas de Markov, a la par que unas notas biográficas sobre el matemático ruso, véase Basharin et al. [2004]; para descripciones más extensas véase Grinstead and Snell [2006, Cap 11], Gamerman y Lopes [2006, Cap 3], Neal [1993, Cap 3].

[7] Los posibles valores de las variables X_i forman un conjunto S, y reciben el nombre de *conjunto de estados de la cadena*, $S = \{s_1, s_2, s_3, ..., s_i, ...s_n\}$.Las transiciones entre los estados sólo pueden producirse entre estados vecinos; por lo tanto, la única forma de alcanzar el estado i es desde el estado $i-1$ o bien el $i+1$

[8] Las probabilidades de transición suelen representarse por el símbolo p_{ij},

[9] En la matriz de transición, se identifica mediante una probabilidad unitaria, $p_{ii} = 1$

1. que desde alguno de sus estados sea imposible la transición a un estado vecino

2. que sea posible ir de un estado cualquiera a un estado absorbente (no necesariamente en un paso).

La aplicación de la imputación múltiple a una cadena de Markov requeire que ésta carezca de estados absorbentes.

Estado estacionario

Cuando en los algoritmos MCMC, aproximemos la función de distribución que genera los datos, utilizaremos la propiedad de estacionariedad de las cadenas de Markov, cuya obtención consiste en el cálculo de una matriz, $\boldsymbol{W}$, que contiene probabilidades invariantes. Formalmente $\boldsymbol{W}$ es el límite de la matriz de transición cuando el número de pasos tiende a infinito,

$$\boldsymbol{W} = \lim_{n\to\infty} \boldsymbol{P}^n = \boldsymbol{P}^\infty$$

El cálculo numérico puede hacerse por diferentes caminos que, obviamente ofrecen el mismo resultado; la elección de uno u otro camino es actualmente irrelevante debido a la enorme potencia computacional disponible. Para calcular $\boldsymbol{W}$ puede optarse por dos vías. La primera consiste en utilizar la propia definición de estacionariedad:[10] buscar un vector, $\boldsymbol{\pi}$, que permanezca inalterado tras multiplicarlo por la matriz de transición $\boldsymbol{P}$; es decir, $\boldsymbol{\pi} \cdot \boldsymbol{P} = \boldsymbol{\pi}$.La segunda vía de cálculo consiste en utilizar al teoría de los valores propios.

1.2.2. Simulación de Montecarlo

La simulación de Montecarlo es un método numérico que permite resolver problemas matemáticos mediante la simulación de variables aleatorias,[11] cuya base teórica era conocida con anterioridad al trabajo Neumann & Ulam [1949], ya que algunos problemas de Estadística se resolvían utilizando muestras aleatorias. Sin embargo, hasta la aparición de la informática, este método no encontraba aplicaciones suficientemente amplias, debido a que ya la simulación a mano de variables aleatorias era muy laboriosa.

Una de las ventajas que ofrece la simulación de Montecarlo es su flexibilidad. Un ejemplo de ello es la realización de la gráfica que representa la frontera eficiente descrita por Markowitz [1952]. A tal efecto, programamos la instrucción `frontera.montecarlo()`,

[10] Dado que el sistema a resolver es homogéneo, para completar las ecuaciones se añade la condición $\sum \boldsymbol{\pi}_i = 1$.

[11] Se considera como fecha de nacimiento del método de Montecarlo el año 1949 con el trabajo de los matemáticos J. von Neumann y S. Ulam (véase Neumann & Ulam [1949]

y la aplicamos a los 5 activos financieros utilizados en el epígrafe *1.1* (el resultado puede verse en las figuras 1.2).

```
frontera.montecarlo ←
function(x,y){

  #Datos iniciales

  var(x)→MCV
  apply(x,2,mean)→Er
  as.matrix(Er)→Er
  N←dim(MCV)[2]

  #Matriz resultados

  sim←matrix(NA,ncol=ncol(x)+2,nrow=y)
  colnames(sim)←c("Risk","Return",colnames(x))

  #Loop Montecarlo

  for (i in 1:y){
    abs(rcauchy(N))→Vt
    as.matrix(Vt)→Vt
    Wt←Vt/sum(Vt)
    Rt←t(Wt)%*%Er
    St←sqrt(t(Wt)%*%MCV%*%Wt)
    sim[i,1]←St
    sim[i,2]←Rt
    sim[i,3:ncol(sim)]←Wt
  }

  #Dibujo Región de Markowitz (1952)

  plot(sim[,1],sim[,2],ylab=expression(mu[c]),xlab=expression(sigma[c]),
  family="mono",
  axes=FALSE,
  col="dodgerblue3")
  axis(side=1,family="Helvetica",cex.axis=0.8)
    axis(side=2,family="Helvetica",cex.axis=0.8)

  #Salida de la instrucción

  sim[order(sim[,1]),]→markowitz; as.matrix(markowitz[1,])→MVP; t(MVP)→MVP;
  rownames(MVP)←c("OptPortf"); return(MVP)
}
```

La sesión de **R** es la siguiente:

```
# Cartera óptima según simulación de Montecarlo
> frontera.montecarlo(activos,50000)

             Riesgo Rendimiento  Amazon  Apple     Ibm  Intel Microsoft
MinVarPortf   1.111     0.04398 0.09043 0.5409 0.02342 0.1008    0.2445
```

A la luz de la figura 1.2 vemos que: (i) mediante la simulación de Montecarlo se resuelve un problema analítico complejo de una forma sencilla y (ii) gracias al elevado número de carteras simuladas es posible representar de forma muy precisa la región de interés.

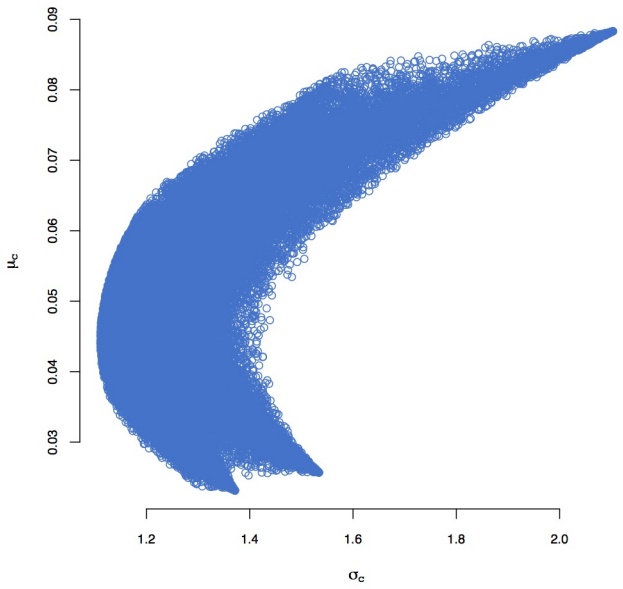

FIGURA 1.2: Región de Markowitz simulando 50000 carteras

Integración de Montecarlo

En un buen número de ocasiones, es necesario evaluar integrales definidas de una función dada:

$$\int_a^b f(x) \cdot dx$$

por el teorema del fundamental del cálculo integral sabemos que se puede encontrar una integral definida $F(x)$, también denominada antiderivada, tal que $\frac{d}{dx}F(x) = f(x)$. Y, según la regla de Barrow,

$$J = \int_a^b f(x) \cdot dx = F(x)|_a^b = F(b) - F(a)$$

La dificultad reside en que muchas funciones $f(x)$ carecen de antiderivada expresada analíticamente o bien, de existir ésta, su obtención resultaría sumamente laboriosa. En tales casos, cabe usar la llamada *integración numérica*, cuya pretensión es aproximar el valor de la integral definida tanto como se desee. Ejemplo típico es, en Estadística, la obtención de la integral Gaussiana,

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} \cdot dx$$

de la que sabemos que $F(0) = 1/2$ y $F(\infty) = 1$

En términos generales, las técnicas tradicionales de integración numérica (como la regla de los trapecios o la regla de Simpson) funcionan muy bien en el caso de unidimensionalidad, pero son poco eficientes en caso de integrales multidimensionales. En cambio las técnicas de simulación de Montecarlo, si bien no ofrecen la máxima precisión en el caso unidimensional, son más eficientes cuando nos encontramos ante dos o más dimensiones.

Aunque existen diversas técnicas de integración mediante simulación de Montecarlo (por ejemplo, una de ellas es el *método hit-and-miss*, que converge muy lentamente) expondremos la más común, referida en la literatura como *integración de Montecarlo.*

Sea $J = \int_a^b f(x) \cdot dx$. La definición de integral de Riemann nos permite escribir,

$$J = \lim_{n\to\infty} \sum_{j=0}^{n-1} h \cdot f(a + j \cdot h)$$

con $h = \frac{b-a}{n}$ y siendo $h \cdot f(a + j \cdot h$ el área del rectángulo de base h y altura $f(a + j \cdot h)$.

Consideremos ahora la variable aleatoria X_n, que toma valores del conjunto $\{a, a + h, \cdots, a + (n-1) \cdot h\}$ con la probabilidad uniforme $\frac{1}{n}$. De acuerdo con esto tendremos que:

$$\begin{aligned} E[f(X_n)] &= \sum_{j=0}^{n-1} f(a+j\cdot h)\cdot P[X_n = a+j\cdot h] \\ E(f(X_n)) &= \frac{1}{n}\cdot\sum_{j=0}^{n-1} f(a+j\cdot h) \\ J &= \lim_{n\to\infty} h\cdot f(a+j\cdot h) \\ &= \lim_{n\to\infty} \frac{b-a}{n}\cdot f(a+j\cdot h) \\ J &= \lim_{n\to\infty} (b-a)\cdot E[f(X_n)] \\ &= (b-a)\cdot \lim_{n\to\infty} E[f(X_n)] \end{aligned}$$

Pero,

$$\lim_{n\to\infty} E[f(X_n)] = E[\lim_{n\to\infty} f(X_n] = E[f(\lim_{n\to\infty} X_n)]$$

Si $X_n \sim U(a,b)$ entonces,

$$E[f(U)] \to J = (b-a)\cdot E[f(U)]$$

Luego, si $U_1, U_2, \cdots, U_n$ son muestras independientes e idénticamente distribuidas de $U(a,b)$ podremos estimar J según la expresión,

$$\begin{aligned} \hat{J} &= \frac{(b-a)}{n}\cdot\sum_{j=0}^{n-1} f(U_j) \\ &= h\cdot\sum_{j=0}^{n-1} f(U_j) \\ \hat{J} &= (b-a)\cdot\widehat{E[f(U)]} \end{aligned}$$

Gracias a esta demostración es muy sencillo incorporar la simulación de Montecarlo a la resolución de integrales definidas, por ejemplo, financieras. En **R** es muy sencillo llevar a cabo la integración de Montecarlo; al efecto hemos programado la instrucción `integral.montecarlo()`,

```
integral.montecarlo ← function(x,a,b,n){

  #variable x: función que se desea integrar
  #variable a: extremo inferior de la integral
  #variable b: extremo superior de la integral
  #variable n: número de simulaciones

    u ← runif(a,b,n)
    x ← sapply(u,f)

    return(mean(x)*(b-a)) }
```

Por ejemplo si se desea integrar la función $log(x)$ entre 1 y 100 procedemos de la siguiente manera:

```
> f←function(x) return (log(x))

#Resultado con 200 simulaciones

> integral.montecarlo(f,1,100,200)
[1] 45.87305

#Resultado con 2000 simulaciones

> integral.montecarlo(f,1,10,2000)
[1] 62.9409

#Resultado con 200000 simulaciones

> integral.montecarlo(f,1,10,200000)
[1] 102.2767
```

1.2.3. MCMC. Algoritmo Gibbs Sampling

MCMC es una colección de métodos diseñados para generar números pseudoaleatorios extraídos de una distribución de probabilidad vía cadenas de Markov.[12] MCMC nace durante la década de los 50 del siglo pasado y su objetivo es resolver los complejos

[12]MCMC está estrechamente relacionado con los paseos aleatorios, característica que crea una división en la clasificación de los algoritmos. Entre éstos se hallan: (i) *Metropolis-Hastings*, que genera un paseo aleatorio usando una función de densidad prefijada; (ii) *Gibbs Sampling*, que como veremos requiere que todas las distribuciones condicionales puedan ser muestreadas (p. 33). Dentro de la categoría de

problemas planteados por la Física de la época, que requerían potentes métodos numéricos, en especial el cálculo numérico de integrales multidimensionales. Afortunadamente, su campo de aplicación no ha permanecido únicamente en la Física, sino que gracias a su desarrollo posterior, se ha aplicado en otros campos, especialmente en Reconstrucción de Imágenes.

En la década de los ochenta del siglo XX, los métodos MCMC empiezan a utilizarse en Estadística, significando una revolución sobre todo para la inferencia bayesiana, donde la información sobre los parámetros desconocidos se expresa bajo la forma de una distribución de probabilidad, cuyo cálculo, incluso en modelos relativamente sencillos, presentaba una carácter muy complejo y a menudo intratable; actualmente, gracias a la introducción de las técnicas MCMC, es posible simular la distribución completa posterior de los parámetros desconocidos. La dinámica de estos algoritmos consiste en construir una cadena de Markov cuya distribución estacionaria coincida con la distribución de interés. La literatura referente a MCMC ha gozado de un auge durante la última década gracias al aumento de la potencia de los ordenadores. Una buena introducción a estos métodos está en Dapugnar [2007, Cap 8] y Jackman [2000]; para descripciones más extensas, ver Gamerman y Lopes [2006] y Neal [1993].

El algoritmo Gibbs Sampling

En 1953, el físico y matemático Nicholas Metropolis y colegas diseñaron un algoritmo para aproximar la función de probabilidad de Boltzmann[13]. El diseño del algoritmo presentaba una novedad importante, la utilización de la cadena de Markov para aproximar la función de distribución; diecisiete años más tarde el algoritmo Metropolis fue generalizado por Hastings, de ahí el nombre compuesto de algoritmo *Metropolis-Hastings*.[14] El trabajo de Hastings fue, sin duda, de una fertilidad extraordinaria, dotando a este algoritmo de una mayor adaptabilidad a distintas situaciones, siendo el algoritmo Gibbs Sampling un buen ejemplo de ello,[15] lo que permite considerar Metropolis-Hastings no como un algoritmo *per se*, sino como una familia de algoritmos, abriendo camino para el diseño de nuevos algoritmos.

El algoritmo Gibbs Sampling,[16] es un algoritmo MCMC y un caso particular del Metropolis-Hastings, creado por Stuart y Donald Geman durante la primera mitad de los años 80 del siglo XX, para su posterior aplicación a la Reconstrucción de Imágenes.[17]

algoritmos que, con la finalidad de mejorar la velocidad de convergencia, evitan los paseos aleatorios está cada vez más difundido el *Hybrid Monte Carlo*, que introduce funciones dinámicas hamiltonianas, Neal [1993].

[13] Véase Metropolis et al. [1953]

[14] Véase Hastings [1970]

[15] Para una aplicación al campo de la visión por computador puede consultarse en Szeliski [1989]; para otras variantes ver: Kennedy & Kuti [1987], Swendsen & Wang [1987] y Goodman & Sokal [1989].

[16] Llamado así como homenaje al físico estadounidense Gibbs.

[17] El trabajo clásico que fija el punto de partida es Geman & Geman [1984]. Otras descripciones pueden

Este algoritmo es conceptualmente el más simple de los que forman parte de MCMC, lo que simplifica su diseño informático.[18] Su fexibilidad ha impulsado el desarrollo de la inferencia bayesiana, gracias al software BUGS (**B**ayesian Inference **U**sing **G**ibbs **S**ampling), incluso se ha propuesto renombrarlo como *Bayesian Sampling*. El Gibbs Sampling parte de la relación de proporcionalidad bayesiana entre las funciones de distribución a priori y a posteriori,

$$p(\theta|x) \propto p(\theta) \cdot p(x|\theta)$$

donde la función de verosimilitud, $p(x|\theta)$, y la distribución a priori, $p(\theta)$, son fácilmente obtenibles, justo lo contrario de lo que sucede con la función a posteriori $p(\theta|x)$, analíticamente inabordable en la mayoría de ocasiones. Sin embargo, $p(\theta|x)$ puede obtenerse utilizando el muestreo dependiente a partir de cadenas de Markov, cuyo estado estacionario sea la función de distribución a posteriori. Así, pues, de forma general, el Gibbs Sampling aproxima la distribución de probabilidad conjunta de dos o más variables aleatorias, mediante el uso de la distribución de probabilidad condicionada de cada variable.

El funcionamiento de este algoritmo es el siguiente: considérese que la distribución de interés es $\pi(\theta)$ donde $\theta = (\theta_1, \theta_2, ..., \theta_d)$; considérese además que disponemos de las distribuciones condicionales de cada variable $\pi_i(\theta_i) = \pi(\theta_i|\theta_{-i})$ para $i = 1, 2, ..., d$. Entonces $\pi(\theta)$ puede calcularse mediante un procedimiento iterativo:

1. Escoger los valores iniciales de $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_d^{(0)})$ en el momento j.

2. Obtener un nuevo valor de $\theta^{(j)}$ desde $\theta^{(j-1)}$ a través de la generación de valores,

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1|\theta_2^{(j-1)}, \cdots, \theta_d^{(j-1)}) \\
\theta_2^{(j)} &\sim \pi(\theta_2|\theta_1^{(j)}, \cdots, \theta_d^{(j-1)}) \\
&\vdots \\
\theta_d^{(j)} &\sim \pi(\theta_d|\theta_1^{(j)}, \cdots, \theta_d^{(j-1)})
\end{aligned}$$

3. Cambiar de j a $j+1$ y volver al paso 2 hasta que se consiga la convergencia.

Si el número de iteraciones ha sido suficiente como para alcanzar la convergencia,entonces el valor de $\theta^{(j)}$ habrá sido extraído de la función de probabilidad $\pi(\theta)$,

consultarse en Schaffer [1997, Cap 3], Neal [1993, Cap 4], Gamerman & Lopes [2006, Cap 5] y Dapugnar [2007, Cap 8]. El contexto de valores NA es discutido por Tanner & Wong [1987]

[18] El Gibbs Sampling es muy fácil de implementar informáticamente pero no lo es de generalizar teóricamente, por lo que es aplicable, en principio, a menos situaciones.

estado estacionario de la cadena de Markov. En el caso de que sea imposible alcanzar la convergencia, pese a un elevado número de iteraciones, puede considerarse que la convergencia lo es **solamente** de forma **aproximada**.

El Gibbs Sampling puede aplicarse en **R** sin dificultad. Para ello hemos diseñado la instrucción `gibbs.sampling()`, que permite obtener una muestra normal bivariante con el coeficiente de correlación lineal que se determine. Si, por ejemplo, queremos que la muestra tenga un ρ de 0.73:

```
gibbs.sampling ← function (n, rho) {

  # variable n: número de simulaciones
  # variable rho: coeficiente de correlación entre las variables

  # construimos la matriz y ajustamos los valores iniciales

    mat ← matrix(ncol = 2, nrow = n)
    x ← 0
    y ← 0

  # iniciamos el loop del Gibbs Sampling

    mat[1, ] ← c(x, y)
    for (i in 2:n) {
      x ← rnorm(1, rho * y, sqrt(1 - rho^2))
      y ← rnorm(1, rho * x, sqrt(1 - rho^2))
    mat[i, ] ← c(x, y)
             }

       return(mat)
}
```

```
#Resultado tras utilizar 50 pasos

> gibbs.sampling(50,0.73) → muestra
> cor(muestra)
          [,1]      [,2]
[1,] 1.0000000 0.6059014
[2,] 0.6059014 1.0000000
```

```
#Resultado tras utilizar 500 pasos

> gibbs.sampling(500,0.73) → muestra
> cor(muestra)
          [,1]      [,2]
[1,] 1.0000000 0.7086282
[2,] 0.7086282 1.0000000

#Resultado tras utilizar 5000 pasos

> gibbs.sampling(5000,0.73) → muestra
> cor(muestra)
          [,1]      [,2]
[1,] 1.0000000 0.7210308
[2,] 0.7210308 1.0000000

#Resultado tras utilizar 50000 pasos

> gibbs.sampling(50000,0.73) → muestra
> cor(muestra)
          [,1]      [,2]
[1,] 1.0000000 0.7307452
[2,] 0.7307452 1.0000000
```

Con esta sesión de **R** vemos que, efectivamente, el algoritmo Gibbs Sampling converge hacia la condición dada a medida aumentamos el número de iteraciones.

Capítulo 2

Imputación múltiple: sección cruzada

"Lo mal conocido se explica por lo bien conocido"
Aristóteles

En este capítulo expondremos las ideas fundamentales de la técnica de imputación múltiple. Para tal fin, haremos un breve recorrido bibliográfico dónde enumeraremos los algoritmos de imputación más relevantes previos a la aparición de dicha técnica, para a continuación explicar el funcionamiento de la imputación múltiple, centrándonos especialmente en la inferencia de Rubin, la cual hemos explorado más a fondo gracias a la simulación de Montecarlo. El capítulo finaliza con un extenso ejemplo basado en el artículo Allison & Chichetti [1976], cuyas dos principales razones que motivan la elección del ejemplo son: presentar la imputación múltiple en el contexto de la sección cruzada, y al mismo tiempo porque la utilización de ésta sirve para enriquecer la investigación realizada por dichos autores.

2.1. Introducción

La imputación múltiple, Rubin [1987], es una técnica estadística basada en simulación de Montecarlo y cadenas de Markov que permite el análisis de bases de datos incompletas, y cuyo avance respecto a los métodos de imputación única consiste en la construcción de un intervalo de confianza que incluya la incertidumbre causada por la presencia de valores NA y, así, evitar la *falsa ilusión de precisión puntual*.[1]

[1] La utilización de imputación única reduce la varianza de los datos, por lo que al construir los intervalos de confianza de los estimadores, éstos resultan más estrechos, consecuencia de la menor variabilidad muestral de los datos imputados.

Para lograr este objetivo, la imputación múltiple se auxilia de algoritmos MCMC, pues éstos consiguen simular valores plausibles mediante la aproximación de la distribución de probabilidad de los datos y generar números pseudoaleatorios gracias a la simulación de Montecarlo. Tras este proceso se consigue $m \geqslant 2$ bases de datos completadas que posteriormente se analizan de forma individual y luego se combinan los resultados utilizando la inferencia de Rubin.

Desde su aparición, la imputación múltiple fue utilizada con entusiasmo por la comunidad científica (en especial por los campos de ciencias de la salud y ciencias sociales) que se enfrentaba a bases de datos incompletas, pues como afirman Heitjan & Rubin [1990], incluso en lo casos más sencillos, la imputación múltiple ofrece mejores resultados que los métodos más sofisticados de imputación única.[2]

Sin embargo, la técnica creada por Rubin no ha estado exenta de críticas, en especial por la utilización de simulación de Montecarlo. Esta objeción ha sido contestada por Schafer, que en distintas ocasiones finalizaba sus exposiciones con el siguiente comentario que citamos textualmente,

> When MI is presented to a new audience, some may view it as a kind of statistical alchemy in which information is somehow invented or created out of nothing. This objection is quite valid for single-imputation methods, which treat imputed values no differently from observed ones. MI, however, is nothing more than a device for representing missing-data uncertainty. Information is not being invented with MI any more than with EM or other well accepted likelihood-based methods, which average over a predictive distribution for the missing data by numerical techniques rather than by simulation.

Las objeciones recibidas por la imputación múltiple han estimulado su avance; por ejemplo Graham et al. [2007] discuten sobre el el número adecuado de imputaciones; Steele et al. [2010] proponen avances en la inferencia de Rubin, pues concluyen que la distribución t de Student no es consistente, y como alternativa consideran el uso de una mezcla de Normales.

2

A nuestro juicio, entre los trabajos más relevantes que utilizan imputación múltiple hallamos: Herzog [1980] y Herzog & Lancaster [1980], que utilizan la imputación múltiple para completar la base de datos de aportaciones individuales de la seguridad social de Estados Unidos; Oh & Scheuren [1980] utilizan esta técnica para estimar el efecto de los NA sobre la varianza en el survey *current population income data*; Burns [1989] utiliza la imputación múltiple para completar la base de datos *Commercial Buildings Energy Consumption System*; Heitjan & Little [1991] aplican la técnica para completar las estadísticas de la *Fatal Accident Reporting System* para la *National Highway Traffic Safety Data*; Kenickell [1991, 1999] aplica esta técnica para imputar los NA del *Survey of Consumer Finance*; Heeringa [1993] completa el *Health and Retirement Survey*; Schenker et al. [1993] imputan los códigos industriales y ocupacionales en bases de datos de uso público;;Khare et al. [1993] y Schafer et al. [1996] analizan el problema de los NA en la *NHANES3*; Barnard & Meng [1999] pasan revista a los estudios de salud en los que se ha aplicado imputación múltiple.

2.2. Algoritmos de imputación previos a la técnica de imputación múltiple

La literatura de imputación se inicia con Yates [1933], cuyo trabajo, enmarcado en un contexto experimental, propone el siguiente algoritmo:

1. **Estimar** la ecuación $\hat{y_i} = \hat{\beta} \cdot x_i$ con los datos disponibles.
2. **Completar** los NA utilizando los valores de las predicciones de la ecuación estimada.
3. **Analizar** la base de datos imputada utilizando las herramientas estadísticas convencionales.

Este algoritmo posee dos inconvenientes: generación de sumas de los cuadrados de los errores demasiado elevadas y, una infraestimación de la matriz de varianzas-covarianzas. Para superar ambas dificultades, Healy & Westmacott [1956] proponen un algoritmo iterativo:

1. **Asignar** a cada NA un valor de prueba.[3]
2. **Estimar** $\hat{\beta}$ mediante mínimos cuadrados ordinarios.
3. **Sustituir** los valores inicialmente desconocidos por las predicciones obtenidas mediante el paso anterior.
4. **Iterar** desde el segundo paso hasta que el vector $\hat{\beta}$ sea invariante según el criterio de convergencia que se establezca.

No obstante, este algoritmo tiene también un inconveniente: la lenta velocidad de convergencia; por ello Pearce [1965] y Preece [1971] proponen mejorarlo mediante aceleradores. Desafortunadamente, este aumento de velocidad puede alterar la suma de los cuadrados de los errores, como indica Jarrett [1978].

Buck [1960] diseña un algoritmo de imputación única también basado en varias regresiones sucesivas de cada variable contra las restantes, que, como indica el propio autor, ofrece resultados aceptables **`si`** los datos se distribuyen mediante una Normal multivariante. Este procedimiento permite hablar de coherencia en el ámbito de la imputación con la ventaja añadida de mejorar la estimación de la matriz de varianzas-covarianzas.[4]

La literatura ha desarrollado también enfoques apoyados en la función de verosimilitud, cuya principal ventaja, como afirman Cox & Hinkley [1974], es la consideración de

[3] El valor de prueba puede proceder de una base de datos anterior, por ejemplo un censo.

[4] Con posterioridad, el algoritmo Gibbs Sampling recoge la idea estimación de una variable contra las demás.

que los parámetros que rigen la distribución poblacional no son fijos, sino que se basan en la verosimilitud de parámetros estimados a partir de la muestra.

Las propiedades de la estimación de máxima verosimilitud fueron aplicadas al campo de los NA con el algoritmo *Expectation - Maximization* (EM), que calcula de forma iterativa la función de verosimilitud, maximizando la expectativa en cada paso.

El origen del alogoritmo *EM* se remonta a McKendrick [1926], siendo Anderson [1957] el primero que lo aplica a bases de datos incompletas; Sundberg [1974] y Beale & Littel [1975] avanzaron en su formulación teórica, pero se centraron en modelos de distribución Normal, y no es hasta Dempster, Laird & Rubin [1977] cuando el algoritmo EM se generaliza y toma como nombre este acrónimo.[5]

El algoritmo EM presenta el inconveniente de su escasa velocidad de convergencia, aunque Little & Rubin [1987] demostraron que el algoritmo EM converge hacia la función de verosimilitud.[6]

2.3. Descripción de la técnica de imputación múltiple

Los métodos básicos de *imputación única* funcionan adecuadamente cuando la presencia de valores NA es muy modesta en número; sin embargo, según Schafer & Graham [2002], no lo hacen cuando el volumen de NA alcanza ratios de órdenes porcentuales superiores al 5 %. Adicionalmente, existen tres problemas relacionados con el proceso de imputación:

1. la elección del algoritmo de imputación;
2. la introducción de posibles sesgos en los resultados debidos al propio algoritmo;
3. negligir la incertidumbre provocada por la presencia de valores NA.[7]

Para superar estos inconvenientes, Rubin [1976, 1978, 1987], replantea el enfoque anteriormente dado al análisis de datos incompletos y, a tal efecto, propone la *imputación múltiple*,[8] cuyo objetivo es *la inclusión de la medida de la incertidumbre provocada por la presencia de valores NA*.[9] Para alcanzar dicho objetivo se reemplaza cada NA por **dos**

[5] El modelo de Healy & Westmacott [1956] es un caso particular del algoritmo EM.

[6] Little & Rubin [1987] estudian también los posibles sesgos del algoritmo según los valores iniciales utilizados.

[7] Los algoritmos de imputación única omiten el componente *between* de la varianza, como veremos en el epígrafe *2.3.2*.

[8] La justificación teórica de la imputación múltiple, según Rubin & Schenker [1987], es más fácil de comprender desde una perspectiva Bayesiana, cuyo desarrollo está disponible en Rubin [1977, 1978, 1979, 1980].

[9] La idea de medir la incertidumbre asociándola al número de NA es una línea de trabajo generada a partir de dos preguntas básicas: ¿Qué estadístico es capaz de establecer esta relación?, ¿Habrá otro estadístico capaz de medir le eficiencia del algoritmo que empleemos?

o más valores, que forman un conjunto de plausibilidad,[10] generando así **múltiples** bases de datos **completadas**.

La imputación múltiple, objeto de estudio de la presente tesis, permite la construcción de un *conjunto de plausibilidad*, para luego combinar los resultados del análisis sobre este conjunto, mediante reglas especiales de inferencia (*inferencia de Rubin*), que permiten medir la incertidumbre relacionando el número de grados de libertad. El siguiente, es un esquema que describe el proceso de utilización de la técnica de imputación múltiple:

1. Etapa de **imputación:** generación de $m \geqslant 2$ bases de datos completadas a partir de métodos MCMC. Los algoritmos MCMC más utilizados son:
   - *Gibbs Sampling* de Geman & Geman [1984].
   - *Data Augmentation* de Tanner & Wong [1987].
2. Etapa de **análisis:** análisis individual de cada una de las m bases de datos generadas en la etapa 1.
3. Etapa de **pooling:** combinación de los m análisis obtenidos siguiendo las reglas inferencia de Rubin [1987].

La descripción de los procesos generadores de valores NA así como la de cada una de las etapas se expone en los tres subepígrafes siguientes.

2.3.1. Análisis de la posible aleatoriedad de los NA

El estudio de la posible aleatoriedad de los NA se realiza a través de la *matriz de respuesta*, que consiste en una recodificación de la base de datos en función de los valores observados. Para ello creamos la variable dicotómica, $\mathcal{R}$, que se define en función del contenido de la celda,

$$\mathcal{R} = \begin{cases} 1 & \text{si la celda contiene un dato observado} \\ 0 & \text{si la celda contiene un NA} \end{cases}$$

de esta matriz se extraen dos informaciones:

Patrón asociado a la *no* respuesta, que identifica regularidades en las posiciones de los NA. En el caso de que las imputaciones se hagan mediante el mismo modelo de no respuesta, entonces la inferencia realizada sobre los m resultados obtenidos representa correctamente la incertidumbre de los valores NA en los resultados del modelo aplicado.

[10] Ver Rubin & Schenker [1986].

Aleatoriedad, que permite descubrir si existe o no aleatoriedad en el conjunto de posiciones de los NA.

Tras la construcción de la matriz de respuesta, se plantea el estudio de las fuentes de pérdida de información asociada a los NA. Para dicho estudio Rubin [1987] propone analizar las relaciones entre $\mathcal{R}$, la parte observada de la muestra para la variable X, que se simboliza por X^{obs}; la parte de la muestra NA de la variable X, que simbolizamos por X^{NA}; y un parámetro representativo de la posible aleatoriedad, ξ. Por consiguiente, deberemos evaluar

$$Pr(\mathcal{R}|X^{obs}, X^{NA}, \xi)$$

Rubin analiza esta expresión de acuerdo con cada una de las fuentes de pérdida de información siguientes:

MCAR (missing completely at random) dada una variable aleatoria X (que contenga NA), se dice que un determinado NA es *completamente aleatorio* si la probabilidad de existencia de éste es independiente de las variables X^{obs} y X^{NA}. Por tanto, dicha probabilidad estará relacionada con el parámetro ξ, es decir,

$$Pr(\mathcal{R}|X^{obs}, X^{NA}, \xi) = Pr(\mathcal{R}|\xi)$$

MAR (missing at random) dada una variable aleatoria X (que contenga NA), se dice que un determinado NA es *aleatorio* si la probabilidad de existencia de éste es independiente de la variable X^{NA}. Formalmente,

$$Pr(\mathcal{R}|X^{obs}, X^{NA}, \xi) = Pr(R|(X^{obs}, \xi))$$

MNAR (missing not at random) dada una variable aleatoria X (que contenga NA), si la existencia de un determinado NA se debe a una causa *conjeturada* como *no aleatoria*, entonces el modelo a utilizar es,

$$Pr(\mathcal{R}|X^{obs}, X^{NA}, \xi) = Pr(\mathcal{R}|(X^{obs}, X^{NA}, \xi))$$

Esta clasificación es importante debido a que la imputación múltiple funciona insesgadamente cuando la fuente generadora de NA es MCAR o MAR (escenarios que utilizamos en esta Tesis).[11]

[11] Si la causa de aparición de valores NA se conjetura como **no aleatoria**, entonces se trata de un proceso más complejo, que ha de analizarse detalladamente para cada base de datos, Rubin[1987, cap. 6], pues de lo contrario pueden aparecer sesgos en el resultado.

2.3.2. Etapas de Imputación y Análisis

Tras las consideraciones precedentes puede iniciarse el proceso de imputación múltiple. La primera parte del proceso es la *etapa de imputación*, durante la cual se genera valores plausibles que reemplazan los NA de la muestra mediante un algoritmo MCMC, cuyo funcionamiento es el siguiente:

1. A partir de la muestra disponible se calcula un vector de medias y una matriz de covarianzas. Estos son los valores iniciales para aproximar la función de distribución.(*cadena de Markov*).

2. Se simulan valores para las celdas vacías escogiendo aleatoriamente valores provenientes de la función de distribución (*Simulación de Monte Carlo*).

3. Se comprueba que el número de iteraciones ha sido suficiente para que la cadena de Markov sea estacionaria. En caso afirmativo, se utilizan las imputaciones finales para crear una base de datos completa; si no hubiera convergencia se ha de volver al paso 1.

Estos nuevos valores simulados que completan la base de datos, forman m muestras completas. En la *etapa de análisis* se utilizan estas m bases de datos completas para estimar un modelo, por lo que cada muestra simulada se trata como una base de datos final.[12]

2.3.3. Etapa de Pooling: inferencia de Rubin

La etapa de pooling resuelve el problema de multiplicidad gracias a la inferencia de Rubin; de la que seguidamente ofrecemos su explicación y fundamentación. Posteriormente intentaremos ir más allá de lo conocido en la literatura al proponer el uso de la simulación de Montecarlo para el tratamiento del problema de multiplicidad.

Se entiende por inferencia de Rubin un conjunto de reglas propuestas por Rubin [1987], cuya finalidad es la construcción de un intervalo de confianza que incluya la incertidumbre (tal y como se puso de relieve en el epígrafe *2.1*). Este cuerpo de reglas descansa sobre la hipótesis de que para un estadístico de interés (referido a X), Q, su media ($\hat{Q}$) y su varianza (U) se pueden expresar como,

$$\begin{aligned} \hat{Q}(X^{obs}, X^{NA}) &\approx E[Q|(X^{obs}, X^{NA})] \\ U(X^{obs}, X^{NA}) &\approx V[Q|(X^{obs}, X^{NA})] \end{aligned}$$

[12] Se calculan tantos modelos como número de imputaciones se haya llevado a cabo.

Además, si la muestra es suficientemente regular[13] y su tamaño adecuado (Rubin [1987]), puede garantizarse un comportamiento asintótico Normal, entonces,

$$(Q-\hat{Q}) \sim N(0,U)$$

Dadas estas dos condiciones, es posible realizar inferencia sobre el estadístico de interés Q. Tras obtener un número m de imputaciones, se calculan m versiones de $\hat{Q}$, simbolizadas por $\hat{Q}_{(j)}$, y otras tantas de U, simbolizadas por $\hat{U}_{(j)}$, con $j = 1,2,\cdots,m$. Por lo tanto,

$$\begin{aligned} \hat{Q}_{(j)} &= \hat{Q}(X^{obs}, X^{NA}_{(j)}) \\ U_{(j)} &= Q(X^{obs}, X^{NA}_{(j)}) \end{aligned}$$

donde $X^{NA}_{(j)}$ simboliza la parte NA *ya imputada* de la muestra. A partir de aquí Rubin [1987] enuncia las siguientes ocho reglas prácticas de inferencia, que conducen a la construcción de un intervalo de confianza para Q y, cuya amplitud depende de m y $B/\bar{U}$:

1. *La estimación puntual combinada de* Q, simbolizada por $\bar{Q}$, que es la media aritmética de las distintas $\hat{Q}_{(j)}$ obtenidas,

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_{(j)}$$

2. *La varianza total asociada a* Q, simbolizada por T, que tiene dos componentes: la varianza *within-imputation*, simbolizada por $\bar{U}$, que es el promedio de las distintas U_j obtenidas,

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U_{(j)}$$

3. y la varianza *between-imputation*, simbolizada por B, calculada mediante

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q}_{(j)}-\bar{Q})^2$$

4. La varianza *total*, definida como,

[13] Rubin emplea el adjetivo *regular* sin precisar su significado. Fruto de la experiencia de los artículos Andreu & Cano (2008) y Cano & Andreu (2010) creemos que podemos dar un significado concreto a este adjetivo: si la muestra es regular las simulaciones generadas mediante MCMC serán plausibles.

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

5. A partir de las varianzas B y $\bar{U}$ se halla el *aumento relativo en la varianza debido a la no respuesta*, representado por r, que se calcula mediante la expresión,

$$r = \left(1 + \frac{1}{m}\right) \frac{B}{\bar{U}}$$

6. Con los estadísticos que se acaban de definir, Rubin [1987] define el intervalo de confianza para Q,[14] cuya expresión es,

$$\frac{(Q - \bar{Q})}{\sqrt{T}} \sim t_{gl,1-\alpha}$$

7. En la anterior expresión, los grados de libertad (gl) se obtienen mediante la fórmula,

$$gl = (m-1)\left(1 + \frac{1}{r}\right)^2$$

   Por lo tanto, r y gl se mueven en direcciones opuestas.

8. El correspondiente intervalo de confianza será,

$$Q = \bar{Q} \pm \sqrt{T} \cdot t_{gl,1-\alpha}$$

Analizando las relaciones de dependencia de los distintos estadísticos introducidos en las reglas de Rubin, llegamos a las conclusión de que el intervalo de confianza de Rubin (ICR) depende en último extremo de X^{obs},$X^{NA}_{(j)}$ y m. Por tanto, dada una muestra X, la variable que gobierna la calidad del ICR descansa en la variable piloto m elegida por el investigador.

2.3.4. Simulación de Monte Carlo sobre la Eficiencia Relativa de Rubin

Como acaba de verse, el número de grados de libertad desempeña un papel fundamental para la determinación de los extremos del intervalo de confianza del escalar Q; sin embargo, dado que gl depende de m y r, es posible establecer una medida de precisión de la imputación múltiple simbolizada por λ, que es la *fracción de información perdida*; basándose en el criterio de información propuesto por Fisher [1935], λ mide la influencia

[14] Según Rubin y Schenker [1986], la inferencia mediante la imputación múltiple es de una elevada precisión.

de la presencia de valores NA sobre el estadístico de interés relacionándose directamente con gl y r,

$$\lambda = \frac{r+2}{r+1} \cdot \frac{1}{gl+3}$$

A partir de λ, Rubin propone una magnitud, a la que denomina *eficiencia relativa* (del algoritmo empleado), RE,

$$RE = \left(1 + \frac{\lambda}{m}\right)^{-1} = \frac{m}{m+\lambda}$$

Esta expresión nos ha permitido construir el cuadro 2.1 que evalúa RE para distintos escenarios. Rubin [1987] y Schafer [1997] deducen, a partir de RE, que un número bajo de imputaciones ($m = 5$) es suficiente para obtener resultados adecuados; así, en Schafer [1997, p. 107] leemos lo siguiente:

> [...] If the fraction of missing information about a scalar estimand is λ, the relative efficiency (on the variance scale) of a point estimate based on m imputations to one based on an infinite number of imputations is approximately $[1+\lambda/m)^{-1}]$ (Rubin 1987, p. 114]. When $\lambda = 0{,}2$, for example, an estimate based on $m = 3$ imputations will tend to have a standard error only $\sqrt{1+0{,}2/3} = 1{,}033$ times as large as the estimate with $m = \infty$. With $\lambda = 0{,}2$, an estimate based on $m = 5$ will tend to have a standard error only $\sqrt{1+0{,}5/5} = 1{,}049$ times as large. In most applications, the additional resources that would be required to create and store more than a few imputations would be not well spent.

La idea expuesta por Schafer, aunque confusa por la inconsistencia entre los cálculos y las fórmulas descritas, es que valores pequeños de m dan lugar a resultados con *muy* poco error y, por consiguiente, utilizar mayor número de imputaciones sería un dispendio de recursos (computacionales). Sin embargo, Schafer [1997, p. 108) matiza que un m reducido no es suficiente para conseguir una estimación robusta de B en el caso multivariante, dado que podría provocar sesgos en el contraste de significación global basado en la F de Fisher. Esa falta de convergencia, debido a un m pequeño, explica que, en este caso, cada vez que se repita el algoritmo de imputación se obtengan resultados sensiblemente distintos.

A efectos de mejorar la comprensión de la inferencia de Rubin y el concepto de eficiencia relativa que de ella se desprende, hemos programado, aprovechando la potencia computacional de **R**, una instrucción que aplica la simulación de Montecarlo sobre las variables independientes de dicha inferencia: m, B y $\bar{U}$, y a partir de éstas, calcula el resto de parámetros. El comando programado lo denominamos `rubin.mc()` y su código es el siguiente,

```
rubin.mc ← function(n){

# Variable n: número de simulaciones.
# Generamos los números aleatorios para m, B y U.

mm ← rpois(n,8)+2
b ← abs(runif(n,0.01,4))
u ← abs(runif(n,0.01,4))

# Aplicamos las reglas de inferencia de Rubin

To ← u+(1+1/mm)*b
bu ← b/u
r ← (1+1/mm)*bu
df ← (mm-1)*(1+1/r)^2
lambda ← (r+2/(df+3))/(r+1)
ARV ← (1+lambda/mm)^(+1)

# Construimos una matriz con los resultados

results ← as.matrix(cbind(mm,bu,r,lambda,ARV))
colnames(results)←c("m","b/u","r","fmi","ARV")
return(results)
}
```

La instrucción `rubin.mc()` además de dibujar las relaciones en la variables elegidas, informa acerca de los valores entre los que se acota cada variable relacionada con la inferencia de Rubin:

```
> inferencia.simulada ← rubin.mc(50000)
      Mean       Min     Q25      Q50      Q75       Max
m    9.994 2.000000 8.0000 10.0000 12.0000  26.0000
b/u 2.990 0.002628 0.5037  1.0019  1.9797 392.2181
r    3.318 0.002878 0.5579  1.1110  2.1976 427.8743
fmi 0.542 0.002871 0.3773  0.5554  0.7189   0.9982
ARV 1.060 1.000206 1.0356  1.0552  1.0782   1.4945
```

A la luz de la figura 2.1 y de la simulación de Montecarlo, cabe efectuar las siguientes afirmaciones:

1. Un elevado número de imputaciones asegura una mayor eficiencia relativa, pero no un descenso de B.

2. Una elevada varianza B causará un r grande, aumentando λ y perjudicando la eficiencia relativa.

3. Fruto del aumento de incertidumbre la varianza total T será elevada, y el intervalo de confianza del estadístico de interés será muy ancho en relación al que se obtendría si la base de datos fuese completa.

4. El parámetro λ es desconocido *a priori*, lo que complica la elección del número adecuado de imputaciones.

5. La fuente de la incertidumbre está en B, por lo que cuanto más diferentes sean las estimaciones puntuales de Q, mayor será B, lo que provocará un intervalo de confianza más ancho debido a la menor precisión en las estimaciones de Q.

Por lo tanto, la idea de un bajo número de imputaciones, aunque comprensible en el momento de la creación de la técnica, creemos que no es aconsejable en la actualidad, dado que se dispone de mayor potencia computacional.

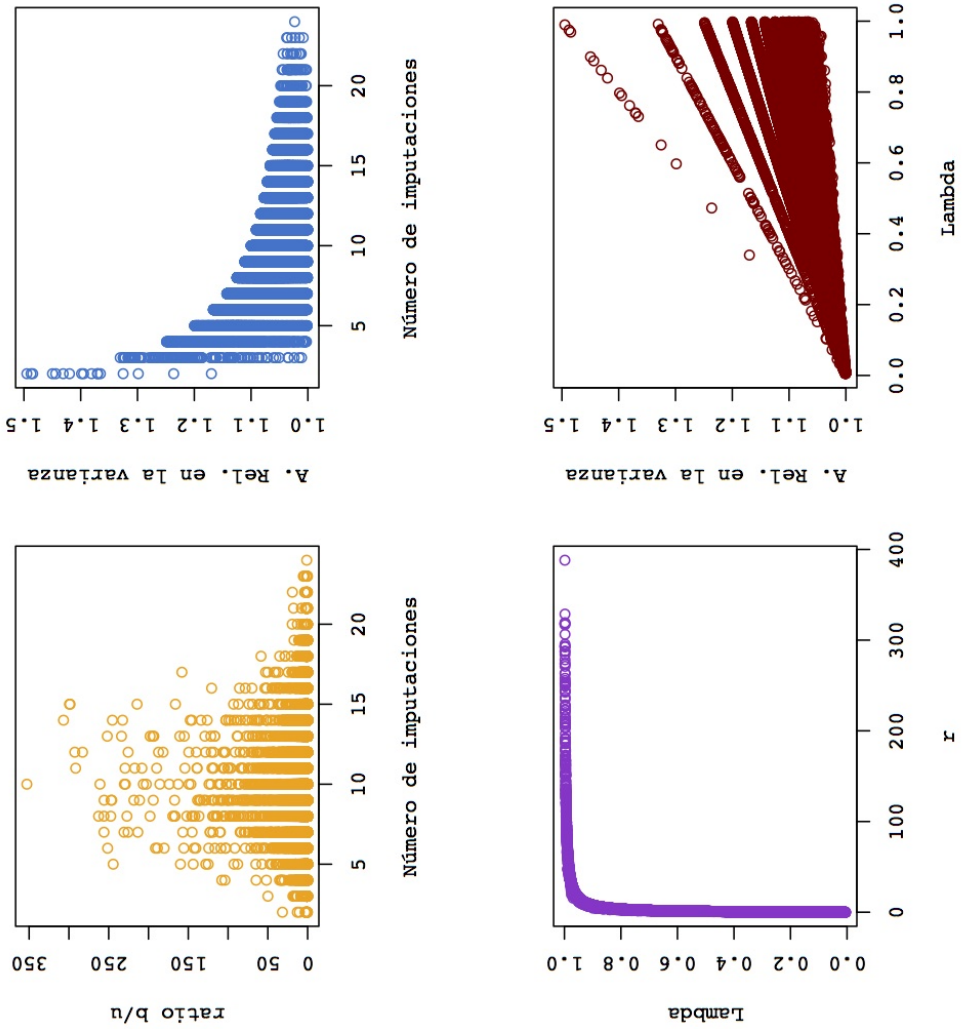

FIGURA 2.1: Relación entre λ y ER tras simular 50000 escenarios

	Fracción de información perdida (λ)									
m	0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
2	0,9756	0,9524	0,9302	0,9091	0,8889	0,8696	0,8511	0,8333	0,8163	0,80
3	0,9836	0,9677	0,9524	0,94	0,9231	0,9091	0,8955	0,8824	0,8696	0,8571
4	0,9877	0,9756	0,9639	0,9524	0,9412	0,9302	0,9195	0,9091	0,8989	0,8889
5	0,990	0,9804	0,9709	0,9615	0,9524	0,9434	0,9346	0,9259	0,917	0,9091
6	0,9917	0,9836	0,9756	0,9677	0,96	0,9524	0,945	0,94	0,9302	0,9231
7	0,993	0,9859	0,9790	0,9722	0,966	0,9589	0,9524	0,9459	0,9396	0,9333
8	0,9938	0,9877	0,982	0,9756	0,970	0,9639	0,9581	0,9524	0,9467	0,9412
9	0,9945	0,9890	0,9836	0,9783	0,9730	0,9677	0,9626	0,9574	0,9524	0,9474
10	0,9950	0,990	0,9852	0,9804	0,9756	0,9709	0,9662	0,9615	0,9569	0,9524
11	0,9955	0,9910	0,9865	0,9821	0,9778	0,9735	0,9692	0,9649	0,9607	0,9565
12	0,9959	0,9917	0,9877	0,9836	0,9796	0,9756	0,972	0,9677	0,9639	0,96
13	0,9962	0,992	0,9886	0,9848	0,9811	0,9774	0,9738	0,9701	0,9665	0,9630
14	0,996	0,993	0,9894	0,9859	0,9825	0,9790	0,9756	0,9722	0,9689	0,966
15	0,9967	0,9934	0,990	0,9868	0,9836	0,9804	0,9772	0,9740	0,9709	0,9677
16	0,9969	0,9938	0,9907	0,9877	0,9846	0,982	0,9786	0,9756	0,9726	0,970
17	0,9971	0,9942	0,9913	0,9884	0,9855	0,9827	0,9798	0,9770	0,9742	0,9714
18	0,9972	0,9945	0,9917	0,9890	0,9863	0,9836	0,9809	0,9783	0,9756	0,9730
19	0,997	0,9948	0,9922	0,9896	0,9870	0,9845	0,9819	0,9794	0,9769	0,9744
20	0,9975	0,9950	0,9926	0,990	0,9877	0,9852	0,9828	0,9804	0,978	0,9756
100	0,9995	0,9990	0,9985	0,998	0,9975	0,9970	0,9965	0,9960	0,996	0,9950
1000	1,0000	0,9999	0,9999	1,000	0,9998	0,9997	0,9997	0,9996	1,000	0,9995
10000	1,0000	1,0000	1,0000	1,000	1,0000	1,0000	1,0000	1,0000	1,000	1,0000

CUADRO 2.1: Eficiencia Relativa de la imputación múltiple

2.4. Aplicación de la técnica utilizando R

Una de las características por las que hemos elegido **R** como instrumento informático es su gran variedad de librerías que implementan la imputación múltiple (ver cuadro 2.4).[15] De entre estas las librerías, hemos decidido usar **mice** desarrollada por Van Buuren & Ousdhoorn [2005, 2009].,[16] porque destaca sobre las otras por ser la más completa en número de instrucciones y por ser la más flexible, pues es posible implementar un modelo de imputación considerando el tipo de variable que se ha de completar y programar la secuencia de visitas del algoritmo.[17] El proceso de imputación múltiple se estructura en 3 instrucciones que se corresponden con las 3 etapas de la técnica:

Etapa	Instrucción	Descripción
Imputación	`mice()`	imputa los datos faltantes y crea un número m de bases de datos completas.
Análisis	`glm.mids()`	analiza los m modelos calculados.
Pooling	`pool()`	pooling de imputaciones mediante inferencia de Rubin.

CUADRO 2.2: La imputación múltiple en **mice**

2.4.1. Dataset Allison & Chichetti [1976]

En el trabajo Allison & Chichetti [1976], los autores investigan las relaciones entre sueño (duración de las fases SWS, *slow wave sleep*, y PS, *paradoxial sleep*), factores ecológicos, y factores anatómicos de 39 especies de mamíferos utilizando una muestra de 62 animales, y manejando las diez variables que figuran en el cuadro 2.3.

Sin perjuicio de que acaso pudieran encontrarse ejemplos igualmente válidos para nuestro propósito, hemos optado por utilizar este dataset debido al siguiente conjunto de características:

- El contenido completo del artículo está disponible en internet.
- El dataset está disponible en *http://lib.stat.cmu.edu/datasets/sleep*.
- El dataset puede ser libremente usado y distribuido con fines no comerciales (cortesía del Dr. Truett Allison).

[15] **R** contiene otras librerías que permiten utilizar métodos de imputación única: **arrayImpute**, **ForImp**, **imputation**, **impute**, **imputeMDR**, **mtsdi**, **missForest**, **robCompositions**, **rrcovNA**, **sbgcop**, **SeqKnn** y **yaImpute**.

[16] Para ver una completa descripción de las instrucciones de esta librería puede utilizarse el sistema de ayuda de **R** y clicar sobre *packages* y ahí seleccionar **mice**.

[17] Traducimos de esta manera la expresión que utiliza **R**, *visit sequence*, procedente de la práctica médica. La expresión hace referencia al orden en el que el algoritmo trata las combinaciones de variables.

Variable	Comentario
BodyWgt	Peso del mamífero (kg).
BrainWgt	Peso del cerebro (g).
NonD (SWS)	Período de tiempo durmiendo sin soñar (horas/día).
Dream (PS)	Período de tiempo durmiendo soñando (horas/día).
Sleep	Período total durmiendo (horas/día).
Span	Esperanza de vida (años).
Gest	Tiempo de gestación (días).
Pred	Índice de predación (rango: 1 *depredador* a 5 *depredado*).
Exp	Índice de exposición durmiendo (rango: 1 *bien protegido* a 5 *más expuesto*).
Danger	Índice de peligro global (rango: 1 *mínimo* a 5 *máximo*).

Cuadro 2.3: Variables utilizadas en Allison & Chichetti [1976]

- El tamaño del dataset es adecuado para realizar tareas computacionales intensivas en un tiempo aceptable.
- El modelo de regresión de sección cruzada que se utiliza es fácilmente replicable.
- La base de datos contiene NA.
- La base de datos está disponible para su utilización directa en **R** con el nombre *sleep* como parte de la librería **VIM**.[18]

Cuando estudiamos la citada investigación, nos dimos cuenta que el empleo de **R**[19] y de la imputación múltiple podía servirnos idealmente (por las características antedichas) para replicar las conclusiones a las que llegaron los autores. Posteriormente, durante el proceso de replicación de resultados, se nos apareció la posibilidad de enriquecer -gracias al uso de esas dos herramientas- los citados resultados.

Para utilizar el dataset Allison & Chichetti [1976] ha de instalarse la librería **VIM** y cargar los datos. Tras disponer de los datos en la memoria, se han de eliminar las filas que contienen NA, lo que conseguimos mediante

```
>install.packages("VIM")
>data(sleep,package="VIM")
```

[18] El dataset se encuentra disponible también en la librería **psy**; sin embargo, en la librería **VIM** es directamente utilizable.

[19] Dada la fecha de publicación del artículo, no cabía la posibilidad de utilización ni de **R** ni de la imputación múltiple.

```
>na.omit(sleep)->sleep.obs
```

Con esta base de datos, Allison & Chichetti [1976] propusieron el siguiente modelo de regresión:[20] $NonD = f(BodyWgt, Danger)$, con la siguiente especificación,

$$NonD = \alpha + \beta 1 \cdot log_{10}(BodyWgt) + \beta 2 \cdot Danger + \epsilon$$

que en **R** lo calculamos de la siguiente forma,

```
>lm(NonD~log(BodyWgt,10]+Danger,data=sleep.obs)->modelo
>summary(modelo)
```

```
Call:
lm(formula = NonD ~ log(BodyWgt, 10] + Danger, data = sleep.obs)

Residuals:
   Min     1Q Median     3Q    Max
-6.379 -1.534 -0.045  2.212  4.747

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.897      0.919   12.95  1.1e-15 ***
log(BodyWgt, 10]   -1.558      0.331   -4.71  3.1e-05 ***
Danger             -0.970      0.318   -3.06    0.004 **
---


Residual standard error: 2.65 on 39 degrees of freedom
Multiple R-squared: 0.545,  Adjusted R-squared: 0.522
F-statistic: 23.4 on 2 and 39 DF,  p-value: 2.15e-07
```

Extraigamos los intervalos de confianza para los regresores,

```
> confint(model)

                    2.5 %     97.5 %
```

[20] Este modelo se estima bajo la hipótesis de homoscedasticidad. Discutiremos esta hipótesis en el epígrafe *2.4.2.*

```
(Intercept)       10.038654 13.7550450
log(BodyWgt, 10] -2.226653 -0.8892081
Danger            -1.612374 -0.3278100
```

Los omisión de valores NA de la bases de datos ha reducido la muestra a 42 individuos, y por eso sólo se tienen 39 grados de libertad. A continuación, estudiaremos este dataset utilizando la técnica de imputación múltiple, para luego proponer un refinamiento estadístico del modelo original.

2.4.2. Aplicación de `mice` al dataset Allison & Chichetti [1976]

En primer lugar describiremos numérica y gráficamente el dataset **sleep**, lo que requiere la instalación y carga de las librerías necesarias.

```
>library(VIM)
>library(mice)
>library(psych)
>data(sleep)

> describe(sleep)

                  var  n   mean      sd median   min     max skew kurtosis
Body.weight         1 62 198.79 899.16   3.34  0.00  6654.0 6.25    40.60
Brain.weight        2 62 283.13 930.28  17.25  0.14  5712.0 4.83    23.24
Slow.wave.sleep     3 48   8.67   3.67   8.35  2.10    17.9 0.28    -0.44
Paradoxical.sleep   4 50   1.97   1.44   1.80  0.00     6.6 1.37     1.78
Total.sleep         5 58  10.53   4.61  10.45  2.60    19.9 0.19    -0.65
Maximum.life.span   6 58  19.88  18.21  15.10  2.00   100.0 1.91     5.00
Gestation.time      7 58 142.35 146.81  79.00 12.00   645.0 1.60     2.32
Predation           8 62   2.87   1.48   3.00  1.00     5.0 0.22    -1.37
Sleep.exposure      9 62   2.42   1.60   2.00  1.00     5.0 0.65    -1.25
Danger             10 62   2.61   1.44   2.00  1.00     5.0 0.36    -1.28
```

Librería	Comentario
Amelia	Crea imputaciones múltiples basado en un modelo Normal multivariante
BaBooN	Genera imputaciones múltiples mediante MCMC. Está enfocado a datos categóricos y fusión de bases datos con el mismo patrón de NA.
cat	Implementa la imputación múltiple de variables categóricas de acuerdo con el modelo log-lineal descrito en Schafer [1997].
Hmisc	Contiene varias funciones para diagnosticar, crear y analizar imputaciones múltiples. La función `fit.mult.impute()` es compatible con la librería **mice**.
kmi	Aplica imputación múltiple utilizando el modelo de Kaplan-Meier, específicamente diseñado para muestras censuradas.
mi	Implementa la imputación múltiple desde una perspectiva bayesiana.
mice	Implementa la imputación múltiple con MCMC utilizando el algoritmo Gibbs Sampling. Es una librería muy completa y flexible, pues permite utilizar matrices que gobiernan la secuencia de imputación. Es la librería que se utiliza en la presente Tesis.
Mimix	implementa un método especial de combinación utilizando una mixtura de Normales.
mitools	Proporciona instrucciones que permiten aplicar la inferencia de Rubin.
MissingDataGUI	Contiene instrucciones para explorar numérica y gráficamente bases de datos incompletas.
miP	Lee y visualiza imputaciones creadas por las librerías **Amelia**, **mi** y **mice**.
mix	Implementa la imputación múltiple para datos mixtos siguiendo el modelo propuesto por Schafer [1997, cap 9]
norm	Implementa la imputación múltiple basado en un modelo Normal multivariante descrito en Schafer [1997, cap. 5 y 6]
pan	Implementa la imputación múltiple en datos de panel.
VIM	Incluye instrucciones para visualizar bases de datos incompletas antes de ser imputadas.
Zelig	Contiene la instrucción `zelig()`, que analiza y combina imputaciones.

Cuadro 2.4: Librerías de Imputación Múltiple en **R**

1. Representación gráfica del dataset resaltando los valores NA sobre el resto de observaciones(figura 2.2).

```
>matrixplot(sleep)
```

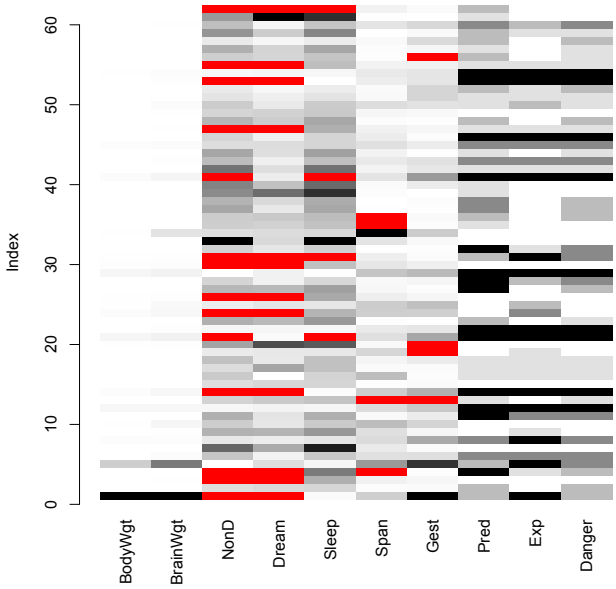

FIGURA 2.2: Mosaico de valores NA en el dataset Allison & Cichetti [1976].

2. Obtención de la matriz de respuesta en función de la variable dicotómica $\mathcal{R}$ usando la función `md.pattern()`,

```
>md.pattern(sleep)

   BodyWgt BrainWgt Pred Exp Danger Sleep Span Gest Dream NonD
42       1        1    1   1      1     1    1    1     1    0
 2       1        1    1   1      1     1    0    1     1    1
 3       1        1    1   1      1     1    1    0     1    1
 9       1        1    1   1      1     1    1    1     0    2
 2       1        1    1   1      1     0    1    1     0    2
 1       1        1    1   1      1     1    0    0     1    2
 2       1        1    1   1      1     0    1    1     0    3
 1       1        1    1   1      1     1    0    1     0    3
         0        0    0   0      0     4    4    4    12   14 38
```

El extremo superior izquierdo de la tabla son los casos completos, y el extremo inferior derecho informa sobre la cantidad de NA presentes en la base de datos, la cual posee 42 casos completos, 20 casos incompletos y 38 valores NA.

3. La matriz $\mathcal{R}$ puede representarse junto a un diagrama de barras que informa acerca de la proporción de NA en las variables de la base de datos; el output de este comando es la figura 2.3.

```
>aggr(sleep, prop=FALSE, numbers=TRUE)
```

4. Extraemos las correlaciones entre las variables que contienen valores NA,

```
>as.data.frame(abs(is.na(sleep)))→ x
>y ← x[which(sd(x)>0)]
>cor(y)

          NonD    Dream    Sleep     Span     Gest
NonD    1.0000  0.90711  0.48626  0.01520 -0.14183
Dream   0.9071  1.00000  0.20370  0.03752 -0.12865
Sleep   0.4863  0.20370  1.00000 -0.06897 -0.06897
Span    0.0152  0.03752 -0.06897  1.00000  0.19828
Gest   -0.1418 -0.12865 -0.06897  0.19828  1.00000
```

Esta tabla informa que las variables de las correlaciones que se establecen entre los NA, por ejemplo la correlación entre NonD y Dream, es muy fuerte, con un 0,91; es decir, tienden a no ser observadas simultáneamente, o que si una es un valor ausente existe una probabilidad elevadísima de que la otra también lo sea. Lo mismo sucede con las variables Sleep y NonD, pero con una correlación menor, dado que en este caso $R = 0{,}49$.

5. También se pueden ver las correlaciones entre las variables que contienen valores NA y el resto de variables completas,

```
>cor(sleep,y,use="pairwise.complete.obs")

              NonD     Dream      Sleep      Span      Gest
BodyWgt    0.22683   0.22259   0.001685 -0.05832 -0.05397
BrainWgt   0.17946   0.16321   0.007859 -0.07921 -0.07333
NonD            NA        NA         NA -0.04315 -0.04553
Dream     -0.18895        NA -0.188952  0.11699  0.22775
Sleep     -0.08023 -0.08023         NA  0.09638  0.03976
Span       0.08336   0.05981   0.005239        NA -0.06527
Gest       0.20239   0.05140   0.159702 -0.17495        NA
Pred       0.04758 -0.06834   0.202463  0.02314 -0.20102
Exp        0.24547   0.12741   0.260773 -0.19292 -0.19292
Danger     0.06528 -0.06725   0.208884 -0.06666 -0.20444
```

Tras explorar el dataset iniciamos el proceso de imputación. Primero la fase de imputación con $m = 5$,

```
>library(mice)
#Etapa de imputación.
>sleep.imp5<-mice(sleep,m=5)
```

y aplicamos el modelo a las bases 5 de datos completadas,

```
#Etapa de análisis.
>sleep.analisis ← glm.mids(NonD~log(BodyWgt,10]+Danger, data=sleep.imp5)
```

Los resultados de cada uno de los m modelos se pueden obtener mediante

```
> sleep.analisis5$analyses

[[1]]

Call:  glm(formula = formula, family = gaussian, data = data.i)

Coefficients:
     (Intercept)  log(BodyWgt, 10]            Danger
         11.3686           -1.1459           -0.8771

Degrees of Freedom: 61 Total (i.e. Null);  59 Residual
Null Deviance:      821.8
Residual Deviance: 499.6  AIC: 313.3

[[2]]

Call:  glm(formula = formula, family = gaussian, data = data.i)

Coefficients:
     (Intercept)  log(BodyWgt, 10]            Danger
         11.3929           -1.1006           -0.7822

Degrees of Freedom: 61 Total (i.e. Null);  59 Residual
Null Deviance:      896.3
Residual Deviance: 616.6  AIC: 326.4

[[3]]

Call:  glm(formula = formula, family = gaussian, data = data.i)

Coefficients:
     (Intercept)  log(BodyWgt, 10]            Danger
         11.6623           -1.1580           -0.8936

Degrees of Freedom: 61 Total (i.e. Null);  59 Residual
Null Deviance:      922.7
Residual Deviance: 591.4  AIC: 323.8

[[4]]

Call:  glm(formula = formula, family = gaussian, data = data.i)

Coefficients:
     (Intercept)  log(BodyWgt, 10]            Danger
          11.576            -0.987            -0.832
```

```
Degrees of Freedom: 61 Total (i.e. Null);  59 Residual
Null Deviance:      882
Residual Deviance: 621.7  AIC: 326.9

[[5]]

Call:  glm(formula = formula, family = gaussian, data = data.i)

Coefficients:
     (Intercept)  log(BodyWgt, 10]            Danger
         11.8731           -1.1710           -0.9923

Degrees of Freedom: 61 Total (i.e. Null);  59 Residual
Null Deviance:      808.9
Residual Deviance: 440.8  AIC: 305.6
```

Puede verse que ahora se han calculado tantos modelos de regresión como bases de datos completadas, y que para cada una de ellas se dispone de coeficientes diferentes. Ahora, para resolver el problema de la multiplicidad se combinan los m resultados mediante la inferencia de Rubin.

```
#Etapa de combinación e inferencia de Rubin.
>sleep.combinado5 ← pool(sleep.analisis5)
>summary(sleep.combinado5)


                    est    se     t    df Pr(>|t|) lo 95  hi 95    fmi
(Intercept)      11.483 0.865 13.27 40.1 4.44e-16   9.73 13.232  0.177
log(BodyWgt, 10] -1.168 0.319 -3.66 44.8 6.68e-04 -1.81 -0.524  0.142
Danger           -0.872 0.312 -2.80 35.5 8.30e-03 -1.50 -0.239  0.212
```

```
#con 50 imputaciones

>sleep.imp50←mice(sleep,m=50]
>sleep.analisis50 ← glm.mids(NonD~log(BodyWgt,10]+Danger,data=sleep.imp50]
>sleep.combinado50 ← pool(sleep.analisis50]
>summary(sleep.combinado50]

                    est    se     t    df Pr(>|t|) lo 95  hi 95    fmi
(Intercept)      11.514 0.873 13.19 47.4 0.000000   9.76 13.271  0.185
log(BodyWgt, 10] -1.137 0.317 -3.59 51.4 0.000749 -1.77 -0.500  0.125
Danger           -0.865 0.315 -2.75 45.2 0.008630 -1.50 -0.231  0.218
```

```
#con 1000 imputaciones

>sleep.imp1000←mice(sleep,m=1000]
>sleep.analisis1000 ← glm.mids(NonD~log(BodyWgt,10]+Danger,data=sleep.imp1000
    ]
>sleep.combinado1000 ← pool(sleep.analisis1000]
>summary(sleep.combinado1000]

                   est    se      t   df Pr(>|t|) lo 95  hi 95    fmi
(Intercept)      11.49 0.855 13.43 50.3 0.000000   9.77 13.204  0.151
log(BodyWgt, 10] -1.13 0.313 -3.60 53.0 0.000695 -1.76 -0.500  0.105
Danger           -0.85 0.306 -2.78 49.1 0.007713 -1.47 -0.235  0.172
```

	Valor	Error Std.	t	gl	Inf 95	Sup 95	λ
α	11.897	0.919	12.95	39	10.038	13.755	-
β_1	-1.558	0.331	-4.71	39	-2.226	-0.889	-
β_2	-0.970	0.318	-3.06	39	-1.612	-0.327	-
			$m = 5$				
α	11.483	0.865	13.27	40.1	9.73	13.232	0.177
β_1	-1.168	0.319	-3.66	44.8	-1.81	-0.524	0.142
β_2	-0.872	0.312	-2.80	35.5	-1.50	-0.239	0.212
			$m = 50$				
α	11.514	0.873	13.19	47.4	9.76	13.271	0.185
β_1	-1.137	0.317	-3.59	51.4	-1.77	-0.500	0.125
β_2	-0.865	0.315	-2.75	45.2	-1.50	-0.231	0.218
			$m = 1000$				
α	11.49	0.855	13.43	50.3	9.77	13.204	0.151
β_1	-1.13	0.313	-3.60	53.0	-1.76	-0.500	0.105
β_2	-0.85	0.306	-2.78	49.1	-1.47	-0.235	0.172

CUADRO 2.5: Resultado de las imputaciones

Comparando los 3 escenarios obtenidos mediante imputación con los resultados ofrecidos en por el artículo original, vemos que: los valores de los coeficientes se han reducido, que se han **recuperado grados de libertad** a medida que aumentaba el número de imputaciones y fruto de esto los intervalos de confianza han cambiado. Desde una perspectiva de Montecarlo, un número grande de imputaciones se corresponde con un número grande de simulaciones, lo que garantiza la convergencia en los valores los parámetros

de la inferencia de Rubin, calculando la verdadera influencia de los valores NA en los resultados.[21]

El contratiempo de utilizar un m elevado es su enorme coste computacional, de ahí el sentido de la *eficiencia relativa* expuesta en la inferencia de Rubin. La variación en los valores de los coeficientes entre cincuenta y mil imputaciones es muy pequeña, a la vez que los cambios en los intervalos de confianza también son reducidos. Por consiguiente, y en consonancia con el cuadro 1.5, cinco imputaciones son suficientes para tener un grado de precisión satisfactorio.

Heteroscedasticidad del modelo de Allison & Cichetti [1976]

Como hemos dicho Alison & Cicchetti [1976] investigaron 39 especies de mamíferos para explicar la cantidad de tiempo en las fases de sueño SWS y PS; en sus resultados, encontraron que la variable que mejor funcionaba en las regresiones era el "peso corporal", sin embargo, nosotros sospechamos que el modelo propuesto por los autores podría estar influido por la presencia de heteroscedasticidad. Para contrastar la hipótesis nula de homoscedasticidad utilizaremos la intrucción `bptest()` de la librería **lmtest**,

```
> library(lmtest)
> bptest(model)

  studentized Breusch-Pagan test

data:  model
BP = 6.2118, df = 2, p-value = 0.04478
```

El resultado del test confirma la presencia de heteroscedasticidad. Para corregir este problema vamos a proponer una especificación alternativa del modelo utilizado en Allison & Cichetti [1976]; en concreto, vamos a estimar la ecuación sin la constante, por lo tanto el modelo a estimar es,

$$NonD = \beta 1 \cdot log_{10}(BodyWgt) + \beta 2 \cdot Danger + \epsilon$$

En **R** lo calculamos de la siguiente forma,

```
>lm(NonD~log(BodyWgt,10]+Danger-1,data=sleep.obs)→model.alt

Call:
lm(formula = NonD ~ log(BodyWgt, 10] + Danger - 1, data = sleep.obs)
```

[21] Los resultados obtenidos con 1000 imputaciones son independientes de los números aleatorios generados para completar las m bases de datos.

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.4721 -0.5805  1.0976  6.5005 12.9120

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
log(BodyWgt, 10]  -2.3719     0.7378  -3.215  0.00258 **
Danger             2.6785     0.3329   8.046 6.86e-10 ***
---


Residual standard error: 6.034 on 40 degrees of freedom
Multiple R-squared: 0.6182, Adjusted R-squared: 0.5991
F-statistic: 32.38 on 2 and 40 DF,  p-value: 4.331e-09
```

Y le aplicamos el mismo contraste para detectar la presencia de heteroscedasticidad,

```
> bptest(model.alt)

  studentized Breusch-Pagan test

data:  model.alt
BP = 2.5108, df = 1, p-value = 0.1131
```

En este caso la hipótesis nula de homoscedasticidad se confirma y por lo tanto los contrastes de significación son válidos. Ahora vamos a estudiar lo que sucedería si utilizásemos el mismo modelo con datos imputados ($m = 1000$),

```
>sleep.analisis1000 ← glm.mids(NonD~log(BodyWgt,10]+Danger-1,data=sleep.
    imp1000]
>sleep.combinado1000 ← pool(sleep.analisis1000]
>summary(sleep.combinado1000]

                     est      se       t     df  Pr(>|t|)  lo 95   hi 95       fmi
log(BodyWgt, 10] -1.472 0.6352 -2.317 57.10 2.411e-02 -2.744 -0.1999
    0.04977
Danger            2.617 0.3122  8.382 57.21 1.559e-11  1.992  3.2424
    0.04787
```

Es destacable el descenso de la magnitud del coeficiente de la variable "peso corporal", aunque mantiene el signo negativo. Sin embargo, lo más impactante de estos resultados es la reducida magnitud de λ, inferior al 5%. La interpretación es que cuando cambia el modelo estimado los valores NA tienen un impacto distinto; comprobemos lo que que sucede si hacemos el mismo análisis utilizando el objeto que contiene 5 imputaciones,

```
>sleep.alt5 ← glm.mids(NonD~log(BodyWgt,10]+Danger-1,data=sleep.imp5)
>sleep.combinado.alt5 ← pool(sleep.alt5)
                     est      se       t     df  Pr(>|t|)  lo 95   hi 95       fmi
log(BodyWgt, 10] -1.462 0.6431 -2.274 55.64 2.685e-02 -2.751 -0.1739  0.06320
Danger            2.612 0.3150  8.292 56.47 2.417e-11  1.981  3.2428  0.05449
```

Estos resultados arrojan luz sobre las afirmaciones de Rubin [1987] y Schafer [1997] relativas al número de imputaciones necesarias. Gracias a la especificación de este modelo λ es muy baja, por lo tanto un elevado número de imputaciones conlleva un aumento de eficiencia relativa muy pequeño, por lo que (en este caso) el aumento de m no compensaría el coste computacional de éste. Sin embargo esta comprobación ha sido posible gracias a la potencia que nos ofrece **R**, pues consigue realizar los cálculos con un coste de tiempo muy aceptable.

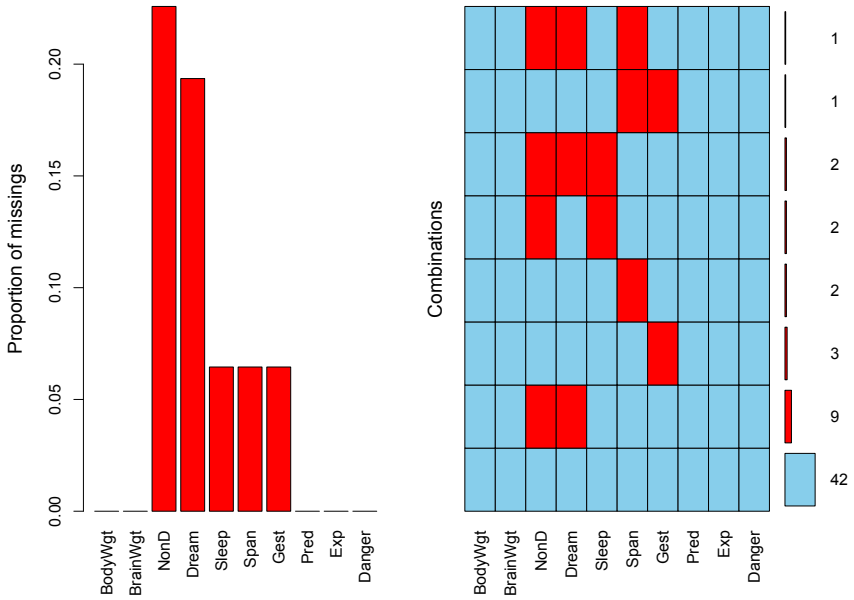

FIGURA 2.3: Representación de la matriz $\mathcal{R}$.

Capítulo 3

Imputación múltiple: series temporales financieras

"El comienzo de la duda es el comienzo de la certidumbre"
E. Ludwig

3.1. Introducción

Andreu & Cano [2008] evaluaron por primera vez la imputación múltiple en series temporales financieras multivariantes de precios bursátiles. El objetivo principal de los autores consistió en ver cómo podía desarrollarse esta técnica en dicho ámbito. Para ello, se utilizaron datos de series temporales financieras de diez empresas presentes en el *Dow Jones Industrial Average* desde 1962 hasta la actualidad, empleándose tres frecuencias distintas: diaria, semanal y mensual.

Los resultados del experimento se evaluaron desde una doble perspectiva, la plausibilidad del valor simulado y el impacto de variaciones en el número de imputaciones, en la proporción de valores NA y en el tamaño de la serie temporal. Los principales resultados de este artículo fueron los siguientes:

Sensibilidad de los errores modificando el porcentaje a imputar: Los resultados de las simulaciones dependen de la proporción de valores NA de la serie temporal. Si el porcentaje de información a completar es del 10 %, los valores simulados son cercanos a a los reales; en cambio, si la proporción de datos a completar es del 50 %, entonces los valores simulados difieren considerablemente de los originales.

Cambios en la longitud de la serie temporal: Los errores crecen a medida que aumenta el tamaño inicial de la serie temporal. Los algoritmos MCMC aproximan la distribución que genera los datos, y ésta cambia a medida que se le añaden más

observaciones. Así aparece, lo que en el trabajo se denominó un *efecto distancia*: a mayor longitud de la serie temporal le corresponde un mayor alisamiento del valor simulado.

Cambios en el número de imputaciones: se produce un ligero descenso del error cuando aumenta m; sin embargo, esta mejora no compensa el elevado coste computacional.

El trabajo Andreu & Cano [2008] puso de manifiesto dos características relacionadas con la utilización de la imputación múltiple en series temporales financieras de carácter bursátil:

1. La matriz de transición de la cadena de Markov asociada a la imputación es fundamental para el correcto uso de algoritmos MCMC.

2. Las series temporales elegidas para acompañar la serie temporal que se desea imputar pueden distorsionar la generación de valores plausibles.

La doble presencia de estas características genera **conjuntos de valores no plausibles**.

En Andreu & Cano [2010] se enfoca a la imputación múltiple en series temporales financieras univariantes. Para ello se diseñó una estructura de datos a partir de retardos de la serie temporal que se deseaba completar. Los resultados de este artículo son mejores que en Andreu & Cano [2008], y se intuye que la estructura de la base de datos influye de forma determinante en el *output* de la imputación múltiple; sin embargo, la no estacionariedad de la series tratadas sigue impidiendo el buen comportamiento del Gibbs Sampling. Este trabajo nos convenció acerca de la importancia del tema de investigación pues tuvimos el honor de ser citados como pioneros por el FMI en "Avoid Filling Swiss Cheese with Whipped Cream: Imputation Techniques and Evaluation Procedures for Cross-Country Time Series" (Denk & Weber [2011]).

Aceptada ya la presencia de heteroscedasticidad, y teniendo en cuenta que en la literatura sobre imputación no hemos encontrado referencias sobre imputación múltiple en STF, el presente capítulo desarrolla un nuevo método que aquí proponemos para llenar este vacío. A tal efecto, para establecer dicho método exponemos sucintamente la teoría de los modelos GARCH, que serán utilizados como filtro. Presentaremos también nuestra librería **`mists`**, que contiene las instrucciones que implementan informáticamente el método propuesto en **R**.

3.2. Modelización GARCH

3.2.1. Justificación y presentación sintética

Muchas series temporales económicas, en particular las financieras habituales de elevada frecuencia, muestran cambios dependientes del tiempo en lo que se refiere a momentos estadísticos condicionados de segundo orden. Como es sabido, estos cambios están sometidos al fenómeno de correlación serial, en el sentido de que ante cambios de gran (pequeña) magnitud en los valores observados, corresponden grandes (pequeños) cambios y periodos de mucha (poca) volatilidad, creando así *clusters de volatilidad* En términos estadísticos, los clusters de volatilidad se traducen en correlaciones positivas de la serie de los cuadrados de los rendimientos logarítmicos, tal como observara Mandelbrot [1963].

El análisis de series temporales financieras con presencia de heteroscedasticidad puso de manifiesto la insuficiencia de los modelos tradicionales hasta el momento, lo que impulsó el desarrollo de modelos aptos para el tratamiento del problema. Engle [1982] propuso el modelo ARCH [1] para solucionar la problemática de la heteroscedasticidad de las series temporales de rendimientos financieros, cuya formulación básica descansa en la ecuación,

$$y_t = \sigma_t \cdot \varepsilon_t$$

donde y_t es la serie temporal financiera que desea modelizarse, σ_t es la volatilidad condicionada asociada a y_t, y ϵ_t es ruido blanco. El siguiente elemento que especifica el modelo ARCH es la distribución condicionada de y_t dado el conjunto de información disponible hasta $t-1$, denotada por Ω_{t-1}, cuya varianza condicionada, σ_t^2, sigue la dinámica

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2$$

donde, para garantizar la positividad de la varianza condicionada, el parámetro ω ha de ser estrictamente positivo y los parámetros α_i, positivos o nulos (pero no todos nulos a la vez).[2] Engle [1982] deduce la condiciones para que el modelo estimado sea débilmente estacionario, y para ello introduce el concepto de *persistencia*, ϕ, que es la tasa a la que el modelo decae.[3] Su expresión matemática es,

[1] El modelo ARCH fue el origen de modelos más sofisticados que tratan de representar el comportamiento de las series temporales financieras de alta frecuencia.

[2] A partir de la distribución de $_t|\Omega_{t-1}$, se construye la función de verosimilitud utilizando la descomposición del error de predicción, permitiendo así obtener estimadores máximo-verosímiles.

[3] A partir de ϕ se obtiene la vida media, $h2l$, que es el promedio (ver, por ejemplo, Ghalanos [2013]) de periodos necesarios para que la volatilidad condicionada alcance la mitad de la incondicionada,

$$h2l = \frac{-ln(2)}{ln(\phi)}$$

$$\phi = \sum_{i=1}^{q} \alpha_i$$

Si se cumple que $0 < \phi < 1$, entonces el modelo obtenido es estacionario; en este caso, la varianza incondicionada es finita y se obtiene mediante,[4]

$$\sigma_y^2 = \frac{\omega}{1-\phi} = \omega \sum_{i=0}^{\infty} \phi^i$$

La utilización del modelo ARCH presenta el problema de, en general, requerir un elevado orden q a fin de que sea posible la captura del comportamiento dinámico de la varianza condicionada. Este problema se debe no tanto a la laboriosidad de manejo del modelo, como al hecho mismo de que puede atentar al principio de parsimonia. Fue Bollerslev [1986] quien, con el fin de simplificar el modelo de Engle, propuso el modelo *ARCH generalizado* o $GARCH(p,q)$, que extiende la especificación de la varianza condicionada del modo siguiente: añade un término autorregresivo a la especificación de la varianza condicionada de Engle, por lo que la especificación GARCH será,

$$\sigma_t^2 = \underbrace{\omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2}_{\sigma_{t(ARCH)}^2} + \underbrace{\sum_{j=1}^{q} \beta_j \sigma_{t-j}^2}_{\text{autorregresivo}}$$

donde $\omega > 0$, $\alpha_i \geqslant 0$ y $\beta_j \geqslant 0$ (pudiéndo suceder que $\beta_j = 0$ para todo j). La persistencia, ϕ, del modelo GARCH es,[5]

$$\phi = \sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j$$

si se cumple que $0 < \phi < 1$, entonces el modelo es estacionario, y la varianza incondicionada se calcula mediante la expresión anteriormente indicada para el modelo ARCH.

En el modelo de Bollerslev [1986], la varianza condicionada depende de los valores históricos de la misma, dando lugar a una una representación más acorde con el principio de parsimonia. De hecho, en la mayoría de aplicaciones empíricas, la especificación $p = q = 1$ es capaz de reproducir con fidelidad la dinámica de la volatilidad de las series

[4] Recuérdese la igualdad $\frac{1}{1-\phi} = 1 + \phi + \phi^2 + \phi^3 + \cdots + \phi^n + \cdots$

[5] En las aplicaciones de modelos GARCH(1,1) a series financieras, es muy común la obtención de una persistencia prácticamente igual a uno, en especial si la frecuencia de observación es alta. Por ejemplo, los trabajos de Engle & Bollerslev [1986], Bollerslev [1987], Baillie & Bollerslev [1989] y Hsieh [1989] con series de tipos de cambio, Chou [1988], Baillie y DeGennaro [1990] y Poon & Taylor [1992] con índices de Bolsa, y otros trabajos citados en Bollerslev, Chou & Kroner [1992] encuentran persistencias superiores a 0.9.

estudiadas.[6] Según Alexander [2009], los parámetros del GARCH pueden interpretarse en términos de la reacción del mercado ante *shocks*:

- El coeficiente α mide la reacción de la varianza condicionada ante noticias del mercado. Cuando α es relativamente elevada (por encima de 0,1), σ_t^2 es muy sensible a eventos del mercado.
- El coeficiente β mide la persistencia de la volatilidad independientemente de lo que suceda en el mercado. Cuando β es relativamente elevada (por encima de 0,9), se necesitan muchos periodos para que la volatilidad se estabilice tras un shock fuerte (por ejemplo una crisis).
- La persistencia ϕ determina la tasa de convergencia de la varianza condicionada hacia σ_y^2. Cuando ϕ es relativamente elevado (por encima de 0,99) indica que los cambios en la varianza condicional son relativamente lentos y, por tanto, los *shocks* en la volatilidad persisten.
- La constante ω, junto a ϕ, determina la magnitud de σ_y^2. Cuando mayor es $\frac{\omega}{1-\phi}$ más elevada es la varianza incondicionada del mercado.

Nelson [1991] observó ciertas limitaciones[7] en el modelo GARCH de Bollerslev [1986]: en primer lugar, las condiciones impuestas sobre los parámetros para asegurar que la varianza condicionada no sea negativa son violadas en algunas aplicaciones empíricas; en segundo lugar, el modelo GARCH es incapaz de modelizar una respuesta asimétrica de la volatilidad ante las subidas y bajadas de los precios. Con el fin de solventar estas deficiencias, la literatura desarrolló modelos GARCH **asimétricos** que permiten que la volatilidad reaccione de forma distinta ante innovaciones de diferente signo. Los GARCH asimétricos más utilizados en la literatura son el GARCH exponencial de Nelson [1991], el GJR GARCH de Glosten et al. [1993] y el Threshold GARCH de Zakoïan [1994][8]

[6] Según Ardia (2008), el GARCH$(1,1)$ se ha convertido actualmente en el "caballo de batalla" tanto de académicos como de analistas financieros.

[7] Las limitaciones originales del modelo GARCH dieron lugar a distintas extensiones del modelo de Bollerslev [1986]. En el cuadro *3.1* mencionamos algunas.

[8] A fin de modelizar el comportamiento asimétrico de y_t, esta tesis utilizará el Threshold GARCH, ver justificación en el epígrafe *3.3.2*

Media incondicionada Bollerslev [1987]	Permite capturar la media de los rendimientos independiente de la variable tiempo (consiste en una constante introducida en la ecuación de la media)
Varianza asimétrica Black [1976] Taylor [1986]	Permite que la varianza condicionada se comporte de forma asimétrica (ésta crece más cuando ocurren innovaciones de signo negativo). Las variantes más usuales en la literatura son: el EGARCH de Nelson [1991], GJR-GARCH de Glosten, Jaganathan & Runkle [1993] y el TGARCH de Zakoïan [1994]
Distr. condicionadas Bollerslev [1987] Nelson [1991]	Utilización de distribuciones de probabilidad con colas más pesadas para $y_t \vert \Omega_{t-1}$: Bollerslev [1987] utiliza la t de Student para modelizar la varianza de rendimientos financieros, y Nelson [1990] introduce la distribución generalizada del error
Curva de noticias Pagan & Schwert [1990] Engle & Ng [1993]	Gráfica que permite ver el impacto de las noticias sobre la varianza condicionada. La ecuación de esta curva varía en función de la especificación elegida; por ejemplo, en el GARCH(1,1) simétrico, la curva de noticias se grafica mediante $\sigma_t^2 = \omega + \beta\sigma_y^2 + \alpha\epsilon_{t-1}^2$.
Rentabilidad-riesgo Engle et al. [1987]	Incluye la volatilidad como variable explicativa de la serie de rendimientos que se desea modelizar. La forma más usual de referirse a este parámetro es el término inglés, *ARCH in mean*.

Cuadro 3.1: Algunas extensiones del modelo GARCH disponibles en la literatura.

3.2.2. Estimación de GARCH en R

La estimación de los modelos GARCH se ha convertido en algo común dentro de los programas econométricos y matemáticos más difundidos, en el que[9] R destaca por tener numerosas librerías dedicadas. En concreto encontramos: **tseries**, que implementa la estimación básica de modelos GARCH; **fGARCH**, desarrollada por el prestigioso equipo *Rmetrics*, con gran variedad de especificaciones para la varianza condicionada (simétricas y asimétricas); **rugarch**, que es la la librería contribuida que más modelos GARCH ofrece, destacando además su sencillez y flexibilidad de uso.[10]

Para llevar a cabo las estimaciones GARCH de la presente de la presente Tesis hemos elegido **rugarch** por las características anteriormente mencionadas. La estimación de un modelo GARCH en R se lleva a cabo mediante dos instrucciones (las instrucciones principales pueden verse en el cuadro 3.2) :

- `ugarchspec()`, donde se especifican ecuación de la media, distribución condicionada de los errores y la dinámica de la varianza condicionada.

- `ugarchfit()`, que estima el modelo que se haya especificado mediante la instrucción `ugarchspec()`.

`ugarchspec()`	Define la especificación de y_t, $y_t\|\Omega_{t-1}$ y σ_t^2 y
`ugarchfit()`	Estima el modelo definido con `ugarchspec()`
`summary()`	Extrae los resultados del modelo estimado
`plot()`	Grafica el modelo, hay doce disponibles
`uncvariance()`	Calcula la varianza incondicionada, σ_y^2
`persistence()`	Calcula la persistencia, ϕ, del modelo estimado
`halflife()`	Calcula la vida media del modelo estimado

CUADRO 3.2: Instrucciones principales de **rugarch** ordenadas en la secuencia lógica de uso.

Para ilustrar la estimación del modelo GARCH(1,1) en R utilizamos la serie diaria de rendimientos del tipo de cambio entre marco alemán y libra esterlina (1984-1993).[11]

[9] Además de en R, los modelos GARCH están disponibles en Stata, SAS, Gretl, Eviews, Matlab y Mathematica entre otros.

[10] R permite también la estimación de modelos GARCH multivariantes mediante **rmgarch** y **gogarch**.

[11] Este dataset ha sido propuesto por McCullough & Renfro [1999] y Brooks et al. [2001] como *benchmark* de prueba del modelo GARCH; la serie se encuentra disponible en R en **bayesGARCH**.

Sobre estos datos estimaremos el GARCH simétrico de Bollerslev [1986],

$$\begin{aligned} y_t &= \sigma_t \cdot \varepsilon_t \\ y_t|\Omega_{t-1} &\sim N(0,\sigma_t) \\ \sigma_t^2 &= \omega + \alpha \cdot \varepsilon_{t-1}^2 + \beta \cdot \sigma_{t-1} \end{aligned}$$

para especificar dicho modelo en **R** nos servimos de `ugarchspec()` y guardamos el output en el objeto **Bollerslev86**,

```
> Bollerslev86 = ugarchspec(mean.model = list(armaOrder = c(0,0),include.mean
    = FALSE, arfima = FALSE), variance.model = list(garchOrder = c(1,1),
   model = "sGARCH"), distribution.model = "norm")
```

A continuación, procedemos a estimarlo con `ugarchfit()` y extraemos los valores de los parámetros del modelo,

```
> ugarchfit(dem2gbp,spec=Bollerslev86) → GARCH
> GARCH@fit$matcoef
         Estimate  Std. Error    t value     Pr(>|t|)
omega  0.01086685 0.002887848  3.762959 1.679148e-04
alpha1 0.15460354 0.026783583  5.772325 7.818530e-09
beta1  0.80442108 0.033856581 23.759666 0.000000e+00
```

Con los coeficientes estimados calculamos directamente la persistencia, la varianza incondicionada y la vida media.

```
> persistence(GARCH)
0.9590246
> uncvariance(GARCH)
0.2652044
> halflife(GARCH)
16.56719
```

Observando los coeficientes del modelo estimado vemos que, de acuerdo con Alexander [2009], la serie de tipo de cambio entre marco alemán y libra esterlina es muy sensible ante noticias en el mercado ($\alpha = 0{,}1546$) mientras que la volatilidad condicionada es poco persistente ($\beta = 0{,}8044$). El bajo valor de la persistencia del modelo ($\phi = 0{,}9590$) indica que se consigue un rápida convergencia ($h2l = 16{,}5671$) hacia la varianza incondicionada de la serie ($\sigma_y^2 = 0{,}2652$).

3.3. Imputación mediante separación: propuesta de un nuevo procedimiento

3.3.1. GARCH bayesianos e imputación múltiple

Tal como ya hemos dicho, la imputación múltiple no puede aplicarse de forma directa a series financieras, debido a que la heteroscedasticidad en varianza impide que el Gibbs Sampling genere valores plausibles. Sin embargo, la aplicación de algoritmos MCMC no es ajena al análisis GARCH de STF univariantes.[12] La literatura ha desarrollado métodos de estimación bayesianos aplicados a los GARCH,[13] basados en MCMC, que pretenden flexibilizar las restricciones sobre los parámetros estimados y realizar inferencia sobre los coeficientes estimados mediante métodos bayesianos.

La estimación bayesiana del modelo GARCH(1,1) está disponible en **R** mediante **bayesGARCH**, que aplicada sobre los mismos datos del ejemplo del epígrafe 3.2 ofrece los siguientes resultados,

```
> library(bayesGARCH)
```

[12] La estimación máximo-verosímil de los modelos GARCH ha recibido algunas críticas por parte de la literatura específica, pues la inferencia sobre los parámetros estimados aún no está resuelta; por ello, existen propuestas alternativas; por ejemplo, Gallant & Tauchen [1989] y Pagan & Schwert [1990] proponen técnicas de inferencia semiparamétricas. Además, la estimación basada en máxima verosimilitud falla cuando el modelo no converge, ya que se han de cumplir las condiciones de estacionariedad para garantizar la positividad de la varianza. Ambos hechos, según Ardia [2008], causan **incertidumbre** sobre los valores estimados de los GARCH.

[13] Ver Geweke [1993], Nakatsuma [1998], Ardia [2007, 2008], Ardia & Hoogerheide [2010]

```
> data(dem2gbp)
> addPriorConditions ← function(psi){psi[2] + psi[3] < 1}
> MCMC ← bayesGARCH(dem2gbp, lambda = 100, delta = 500,control = list(n.chain
    = 2, l.chain = 2000, addPriorConditions = addPriorConditions))
```

```
> summary(MCMC)

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

            Mean      SD  Naive SE Time-series SE
alpha0   0.01147 0.00269 4.253e-05      0.0002959
alpha1   0.16130 0.02598 4.108e-04      0.0028858
beta     0.79591 0.03123 4.937e-04      0.0039066
```

Como se aprecia en estos resultados, los coeficientes estimados mediante métodos bayesianos son muy similares a los obtenidos mediante máxima verosimilitud. Este ejemplo motiva la suposición de que existe un conexión entre MCMC y las STF y *ello justifica que la imputación múltiple sea aplicable a series financieras.* Para implementar la imputación múltiple es necesario diseñar una estrategia que permita la generación de valores plausibles. La plausibilidad exige que los valores simulados satisfagan las dos siguientes condiciones:

- carácter asimétrico de la volatilidad (Black [1976], Taylor [1986])
- presencia de colas pesadas (Mandelbrot [1963])

La búsqueda de la satisfacción de ambas condiciones nos conduce a, en lugar de tratar una STF (y_t) como un todo, separarla mediante un filtro GARCH *asimétrico*, que la descompone multiplicativamente en dos procesos:

- **volatilidad** (σ_t), para capturar la asimetría
- **innovaciones** (ε_t), para capturar la leptocurtosis

Esta es, precisamente, la idea clave que proponemos: *escindir la serie original en dos partes, para posteriormente generar una serie temporal imputada y plausible.*

Intentando desarrollar esta idea estudiamos distintas herramientas analíticas ya presentes en la literatura. Como se indicó en la *Introducción*, encontramos que el Threshold GARCH, y los algoritmos ABB y Gibbs Sampling podían ser utilizados para la generación de valores imputados plausibles. La citada idea clave, y el uso de las herramientas

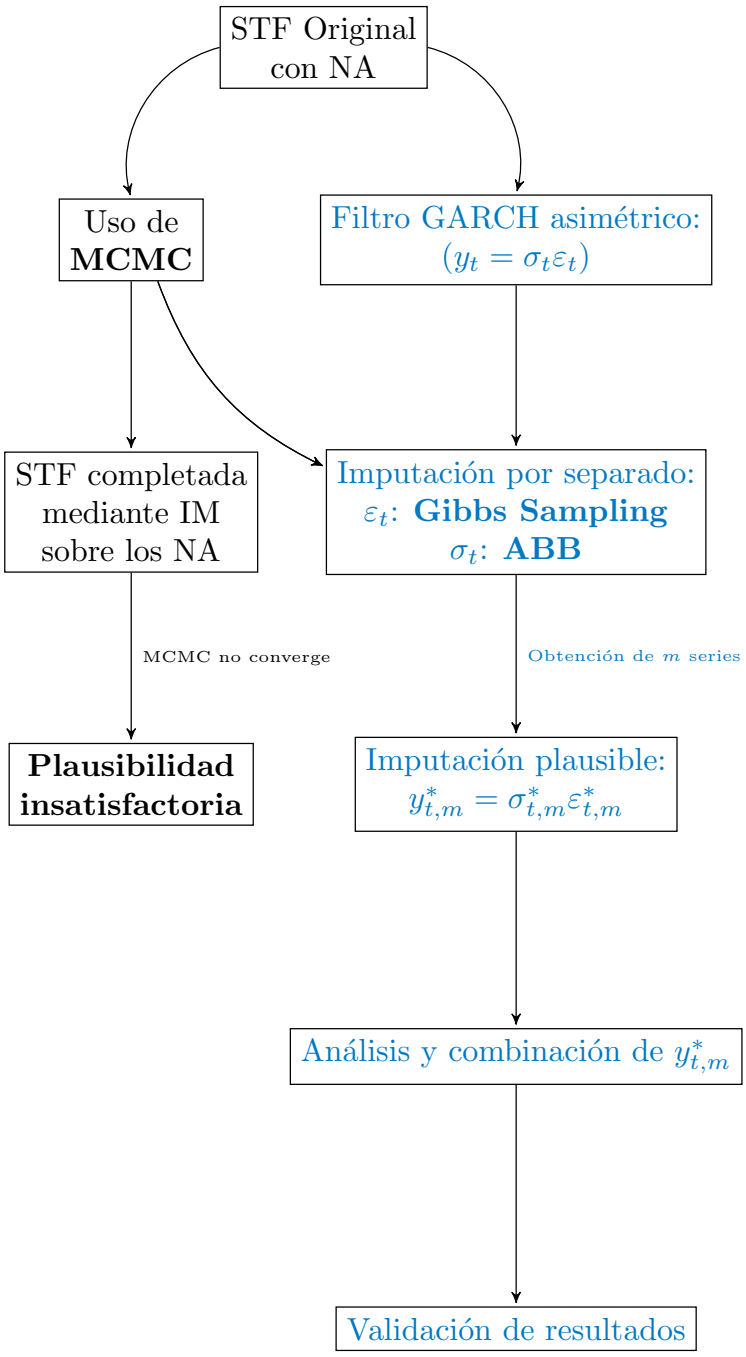

FIGURA 3.1: Esquema de funcionamiento del procedimiento que proponemos (en negro se indica el proceso cuando se trata y_t como un todo y azul cuando se escinde multiplicativamente).

analíticas indicadas configura un nuevo procedimiento al que llamamos método de **imputación mediante separación**. El esquema secuencial de la implementación del método se muestra en la figura *3.1*, donde se distingue la parte izquierda, que corresponde al tratamiento de la serie como un todo, de la parte derecha, que refleja el nuevo método.

La separación de y_t en σ_t y ε_t, permite tratar estas componentes diferentemente: *Gibbs Sampling* para imputar las innovaciones (son estacionarias en media y varianza), y *Approximate Bayesian Bootstrap* para imputar la volatilidad (de distribución desconocida).[14]

Tras imputar aisladamente σ_t y ε_t, efectuaremos su producto, lo que dará lugar a una STF imputada y *presumiblemente plausible*, que simbolizamos por y_t^*.[15]

El resto del capítulo sigue la pauta reflejada en los tres primeros bloques de la parte derecha en la figura *3.1*, comenzando con la exposición de las herramientas analíticas que el método de separación requiere.

La aplicación concreta del nuevo procedimiento a un conjunto de STF previamente elegidas la realizaremos en el capítulo 4, donde nos valdremos de un artificio: tomaremos un conjunto de STF completas y generaremos diversos escenarios en forma de porcentajes de NA, trabajando a continuación sobre dichos escenarios. Luego, desarrollaremos las etapas representadas por los dos últimos bloques de la parte derecha de la figura *3.1*, que son análisis y combinación, y finalmente realizaremos la validación de resultados.

[14] Las figuras *3.2* y *3.3* ilustran las densidades de σ_t y ε_t para General Electric, en las que queremos destacar la forma irregular de la densidad de la volatilidad.

[15] La eventual plausibilidad de y_t^* se evaluará en el capítulo *4*.

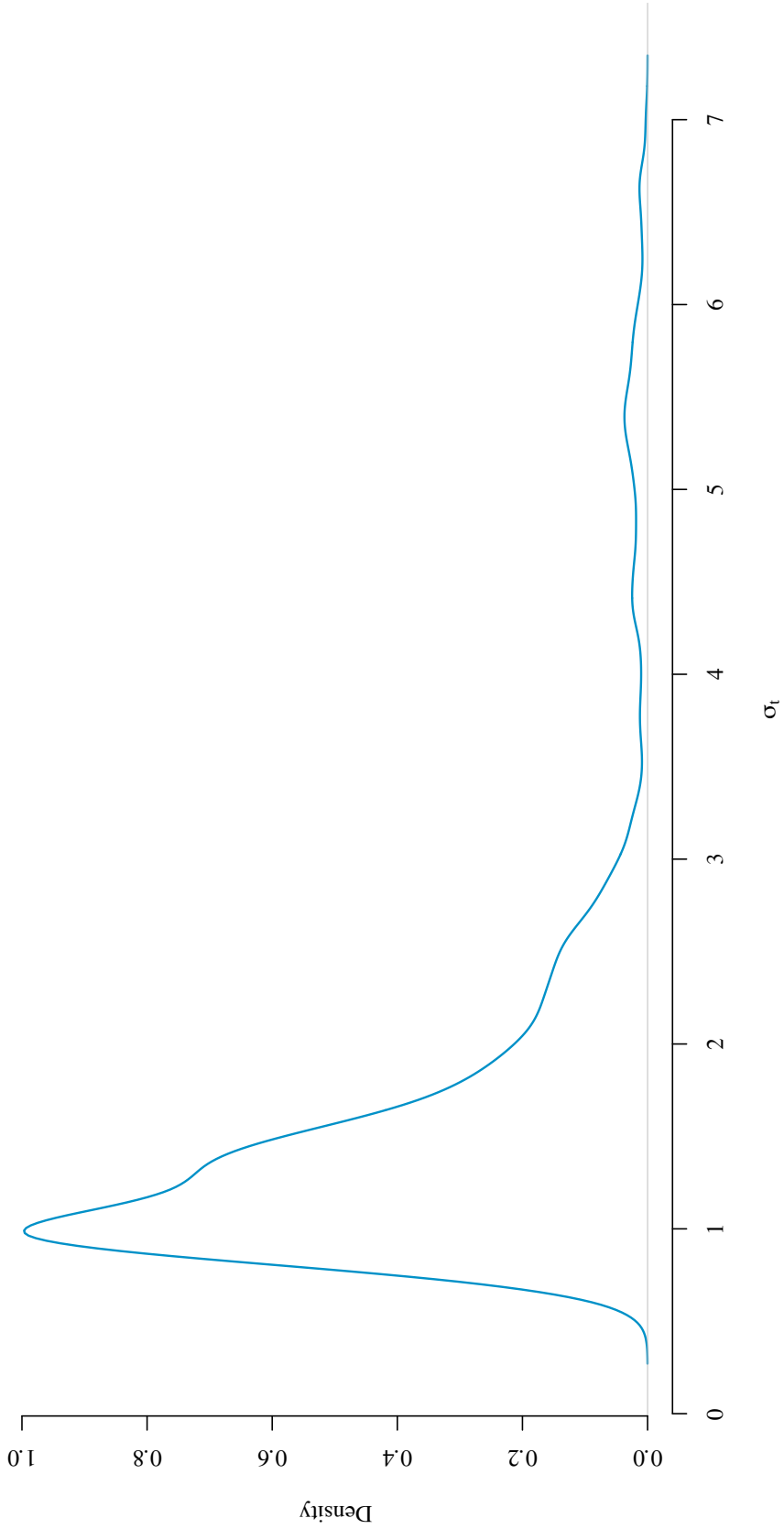

FIGURA 3.2: Densidad de la volatilidad (σ_t) para General Electric durante 2003 - 2012.

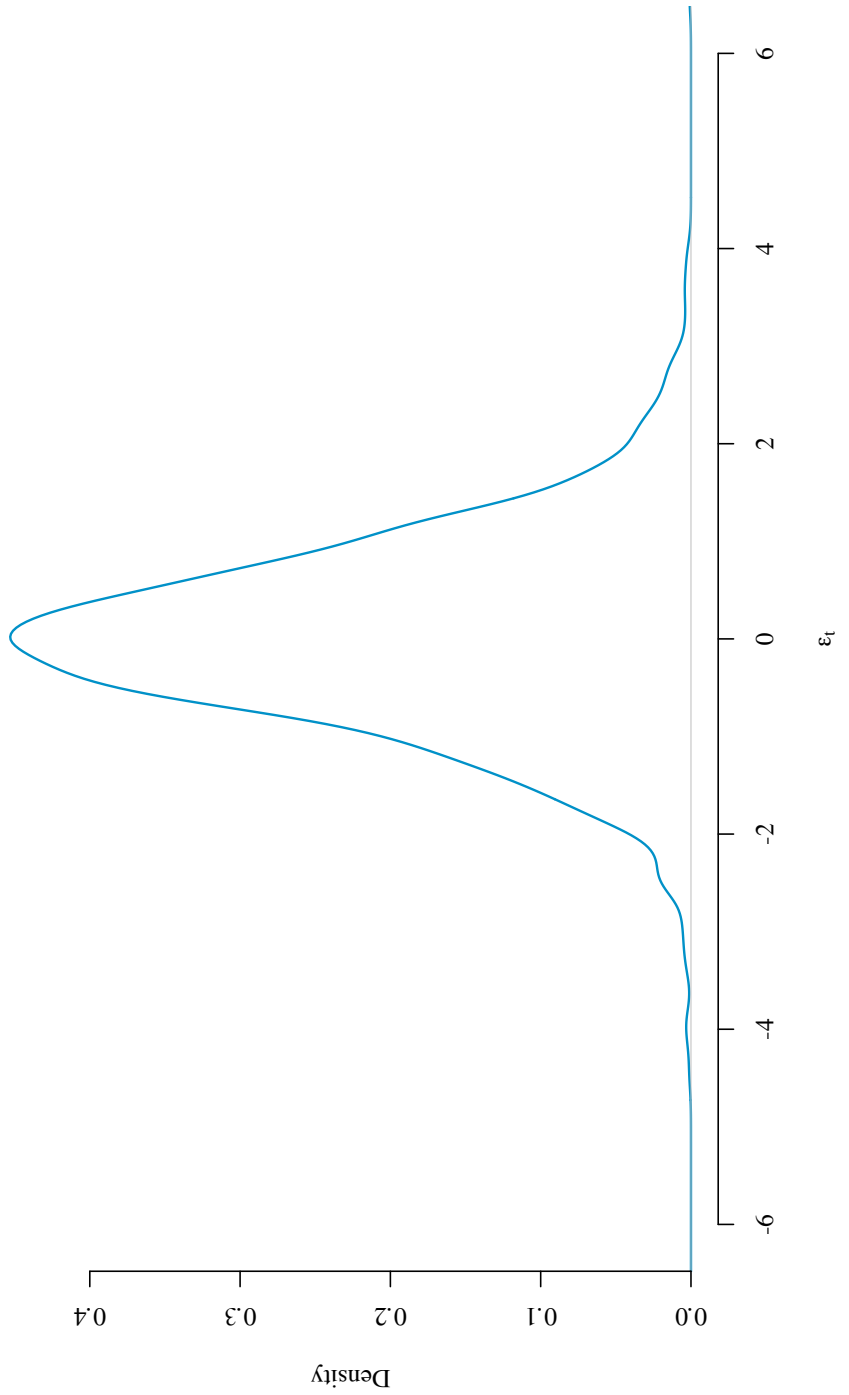

FIGURA 3.3: Densidad de las innovaciones (ε_t) para General Electric durante 2003 - 2012.

3.3.2. Filtrado mediante GARCH asimétrico

Por razones que justificaremos algo más abajo, el GARCH asimétrico que utilizamos como filtro es el *Threshold* GARCH de Zakoïan [1994]. Este modelo amplía el Taylor-Schwert GARCH introduciendo un término asimétrico, permitiendo así que la **desviación estándar condicionada**[16] reaccione de forma distinta ante innovaciones de diferente signo. La dinámica del modelo (con $p = q = 1$) es,

$$\sigma_t = \underbrace{\omega + \alpha|\varepsilon_{t-1}| + \beta\sigma_{t-1}}_{\text{Taylor-Schwert}} + \underbrace{\tau \cdot I(\varepsilon_{t-1} < 0)|\varepsilon_{t-1}|}_{\text{término asimétrico}}$$

donde: $\omega > 0$, $\alpha \geqslant 0$, $\beta \geqslant 0$, y $\tau \in [-1, 1]$; $I(\cdot)$ es una función dicotómica que depende del signo de ε_{t-1},

$$I = \begin{cases} 1 & \text{si } \varepsilon_{t-1} < 0 \\ 0 & \text{si } \varepsilon_{t-1} \geqslant 0 \end{cases}$$

El carácter asimétrico puede verse en la curva de impacto de las noticias (NIC), que se define en función del signo de las innovaciones,

$$NIC = \begin{cases} \omega + \beta\sigma_y + (\alpha + \tau) \cdot |\varepsilon_{t-1}| & \text{si } \varepsilon_{t-1} < 0 \\ \omega + \beta\sigma_y + \alpha|\epsilon_{t-1}| & \text{si } \varepsilon_{t-1} \geqslant 0 \end{cases}$$

Podemos aducir ahora los motivos que justifican la elección del *Threshold* GARCH como filtro:

Asimetría, que permite que la volatilidad condicionada aumente en presencia de innovaciones de signo negativo.

Eficiencia, según Davidian & Carroll [1987] y Zakoïan [1994] es un modelo muy eficiente cuando la distribución de $y_t|\Omega_{t-1}$ no es gaussiana (en esta Tesis utilizamos la GED de Nelson [1991])

Robustez, al tratarse de una especificación basada en valor absoluto, adquiere la condición de robusta ante la presencia de *outliers*; de hecho, Hentschel [1995, p.75] nos dice:

[16] En la literatura, no hay consenso sobre si es preferible modelizar la varianza condicionada o la desviación estándar condicionada. En este sentido, el propio Zakoïan [1994, p. 933] nos dice:

> We adopt a somewhat different approach, based on a paper from Davidian and Carroll (1987) about variance function estimation. One of the most interesting results that they obtain (though variance is allowed to depend on directly observable variables) is that in the case of nonnormal distributions, absolute residuals yield more efficient variance estimates than squared residuals. Therefore we do not square the positive and negative parts of the noise, and we specify the conditional standard deviation instead of the conditional variance.

> [...] the absolute value GARCH model is a more efficient filter of the conditional variance than Bollerslev's (1986) GARCH.

3.3.3. Generación de las innovaciones (ϵ_t) y volatilidad (σ_t)

Las innovaciones son un proceso estacionario en media y varianza, por lo que podemos utilizar el Gibbs Sampling para generarlas. La construcción del Gibbs Sampling para ε_t sigue el esquema: considérese que la distribución de interés es $\pi(\varepsilon)$ donde $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_d)$; considérese además que disponemos de las distribuciones condicionales de cada variable $\pi_i(\varepsilon_i) = \pi(\varepsilon_i|\varepsilon_{-i})$ para $i = 1, 2, ..., d$. A partir de aquí iniciamos el bucle que converge hacia $\pi(\varepsilon)$:

1. Escogemos los valores iniciales de $\varepsilon^{(0)} = (\varepsilon_1^{(0)}, \varepsilon_2^{(0)}, ..., \varepsilon_d^{(0)})$ en el momento j.

2. Obtenemos un nuevo valor de $\varepsilon^{(j)}$ desde $\varepsilon^{(j-1)}$ a través de la generación de valores,

$$\begin{aligned} \varepsilon_1^{(j)} &\sim \pi(\varepsilon_1|\varepsilon_2^{(j-1)}, \cdots, \varepsilon_d^{(j-1)}) \\ \varepsilon_2^{(j)} &\sim \pi(\varepsilon_2|\varepsilon_1^{(j)}, \cdots, \varepsilon_d^{(j-1)}) \\ &\vdots \\ \varepsilon_d^{(j)} &\sim \pi(\varepsilon_d|\varepsilon_1^{(j)}, \cdots, \varepsilon_d^{(j-1)}) \end{aligned}$$

3. Cambiamos de j a $j+1$ y volver al paso 2 hasta que se consiga la convergencia.

Si el número de iteraciones ha sido suficiente como para alcanzar la convergencia, entonces el valor de $\varepsilon^{(j)}$ se extrae de la función de probabilidad $\pi(\varepsilon)$, estado de equilibrio de la cadena de Markov. En la figura *3.4* vemos cómo las funciones de distribución de ε_t imputadas para General Electric (tras eliminar aleatoriamente el 20 % de la serie) se aproximan adecuadamente a la función de distribución real.

* * *

La naturaleza del proceso de volatilidad es diferente al de ε_t, ya que su función de distribución de probabilidad es desconocida. Para tratar este tipo de procesos tuvimos en cuenta la idoneidad de los métodos Bootstrap, propuestos originalmente por Efron [1979], debido a que no presuponen forma alguna a la función de distribución (contrariamente al Gibbs Sampling, que la busca porque presupone su existencia).El bootstrap de Efron [1979] obtiene la función de probabilidad F a partir de una muestra Z_i utilizando la expresión,

$$F_n = \sum_{i=1}^{n} \omega_i \cdot \delta_{Z_i}$$

dónde δ_{Z_i} es una probabilidad asociada a Z_i, ω_i es la ponderación de Z_i, $\omega_i \geqslant 0$, $\sum \omega_i = 1$. Rubin [1981] propone la variante bayesiana del bootstrap de Efron [1979], llamado *Bayesian Bootstrap*,considerando las ponderaciones ω_i como parámetros desconocidos. Para tal fin, Rubin [1981] utiliza la *prior* no informativa $\prod_{i=1}^{n} \omega_i^{-1}$ para obtener el vector ω. Esta *prior*, combinada con la probabilidad multinomial de Z, converge a la distribución de $Dirichlet(1, ..., 1)$, a partir de la cual puede utilizarse simulación de Montecarlo.

Rubin & Schenker [1986] proponen la variante *Approximate Bayesian Bootstrap* adaptada al contexto de la imputación múltiple, que tiene la ventaja de utilizar menos recursos computacionales que el *Bayesian Bootstrap*. La diferencia entre ambos métodos radica en que el *ABB* utiliza la distribución *scaled multinomial* para extraer ω, en lugar de utilizar la distribución de Dirichlet.[17] En la figura *3.5* vemos como el algoritmo ABB consigue generar valores plausibles para σ_t tras eliminar aleatoriamente el 20 % de la serie de General Electric. Las densidades de las series imputadas coinciden de forma muy aproximada con la función de distribución real.

3.4. `mists`: Librería de R que hemos contribuido para implementar el método propuesto

3.4.1. Presentación de la librería

Para la presentación que aquí hacemos de la librería, hemos decidido que resulta preferible, en primer lugar, describir sucintamente las instrucciones que utilizamos en la Tesis y luego a modo de ilustración, la aplicaremos en un ejemplo concreto.

Estas instrucciones forman parte de un cuerpo más amplio, así como de un Manual. Para que este manual pueda ser generado, es necesario que la librería en su conjunto (instrucciones y documentación) pase exitosamente una **exhaustiva** colección de tests por parte de del propio **R**. Una vez superados los tests y generado el *Manual*, estaremos en disposición de afirmar que nuestra librería cumple con los requisitos de R para ser publicada. Finalizamos la presentación de **`mists`** con un epígrafe que contiene el código de las instrucciones, directamente utilizables, que implementan el método de imputación mediante separación.

Uno de los retos que tiene la presente Tesis es programar en **R**, con nuestros propios esfuerzos, las instrucciones que implementan el método propuesto en el epígrafe *3.3*, pues no existe otra librería de **R** capaz de tratar, mediante imputación múltiple, una serie

[17] Rubin & Schenker [1986] demuestran que los métodos *ABB* y *BB* tienen el mismo vector de medias y la misma matriz de correlaciones; sin embargo, la varianza del *ABB* es ligeramente mayor.

temporal financiera. La librería resultante la hemos denominado *múltiple imputation simulation for time series* (**mists**). Esta librería se auxilia de otros paquetes ya disponibles R:

- **rugarch**, estima los modelos GARCH utilizados.
- **LaplacesDemon**, implementa el *Approximate Bayesian Bootstrap*.
- **mice**, implementa el *Gibbs Sampling* para realizar imputación múltiple.

La librería **mists** evalúa la imputación de una STF descomponiéndola en otras dos, utilizando para ello la simulación de valores NA. La columna vertebral de **mists** la configuran las siguientes instrucciones:

- `Filter`: Descompone y_t en σ_t y ε_t mediante el Threshold GARCH.
- `Make.NA()`: Introduce **aleatoriamente** el porcentaje deseado de valores NA en las series σ_t y ε_t (los NA ocupan la misma posición en ambas series).
- `Impute()`: Imputa m veces σ_t (ABB) y ε_t (*Gibbs Sampling*), y crea $y^*_{t,m}$.
- `Garch.mids()`: Estima m modelos GARCH especificado para $y^*_{t,m}$.
- `Rubin.global()`: Calcula la inferencia de Rubin para cada coeficiente estimado del modelo GARCH especificado en `Garch.mids()`.

3.4.2. Ejemplo de uso

A fin de ilustrar cómo funciona la imputación mediante separación utilizamos un ejemplo sobre la serie diaria de rendimientos de Apple (2003-2012). Primero filtramos la serie original para dividirla en volatilidad e innovaciones

```
> x.f ← Filter(APPLE)
```

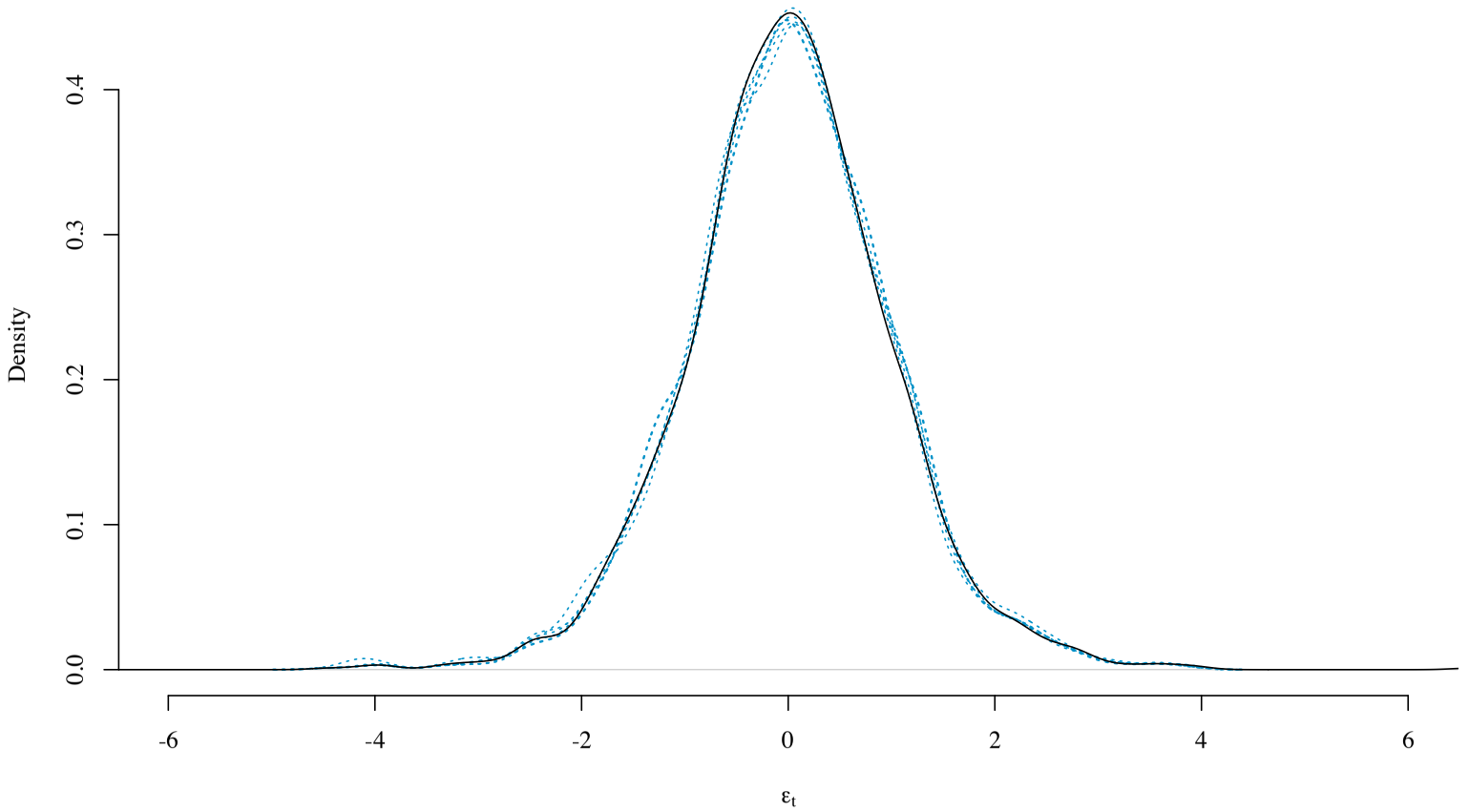

FIGURA 3.4: Comparación de las densidades de ε_t (negro para los datos reales, azul para las imputaciones) para General Electric durante 2003 - 2012. Porcentaje de NA=20 %

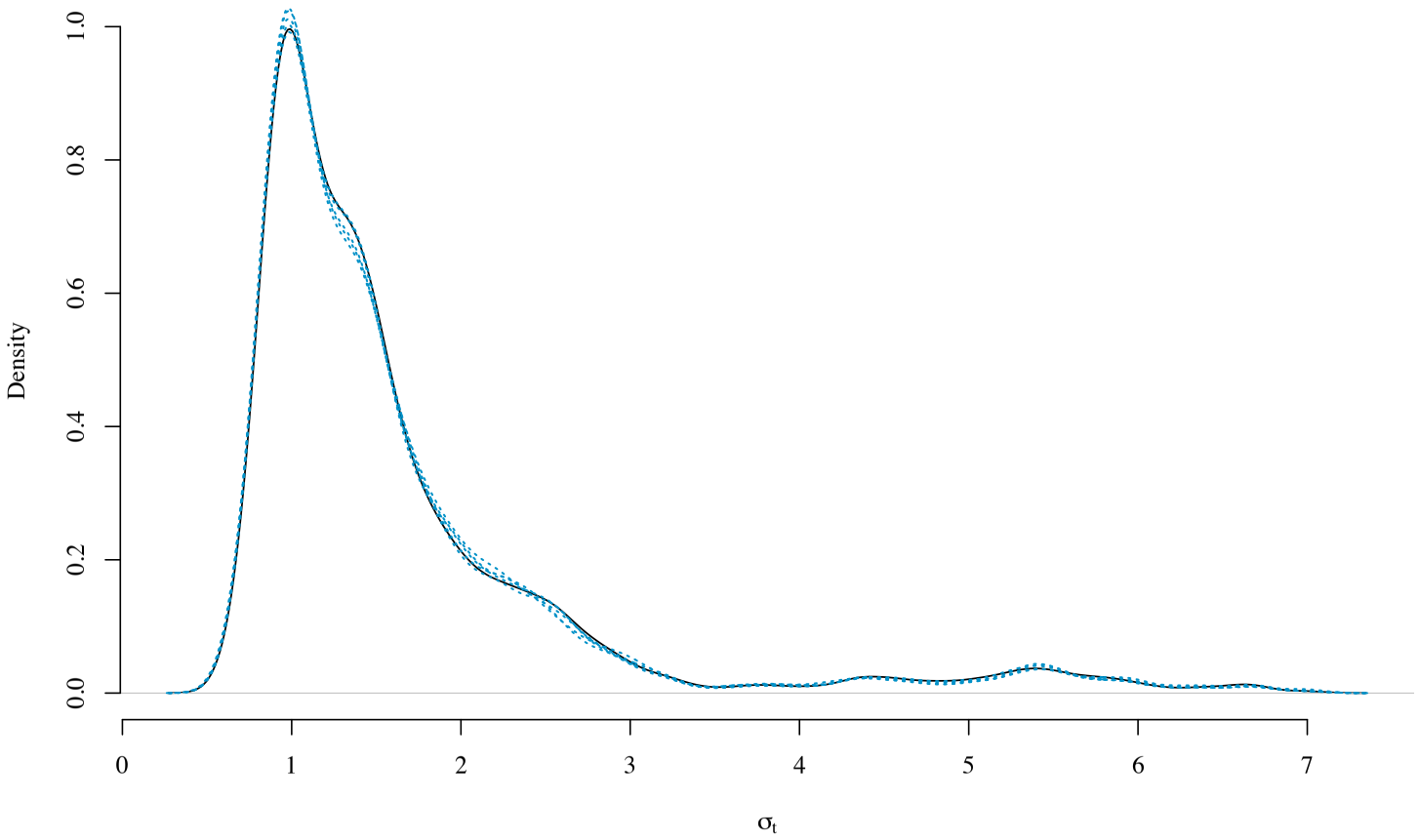

FIGURA 3.5: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para General Electric durante 2003 - 2012. Porcentaje de NA=20 %

En segundo lugar, introducimos un 20 % de valores NA en la serie filtrada,

```
> x.mis ← Make.NA(x.f$M,p=0.2)
```

A continuación, imputamos σ_t y ε_t y creamos y_t^*

```
> x.imp ← Impute(x.mis$Datum)
```

En cuarto lugar, evaluamos las series utilizando el TGARCH de Zakoïan [1994],

```
> analysis ← Garch.mids(x.imp,y=Zakoian(),z="solnp")
```

Por último, aplicamos la inferencia de Rubin sobre los m modelos obtenidos,

```
> Rubin.global(analysis) → RUBIN.Apple
```

```
# Coeficientes estimados mediante imputación múltiple
              Q       SE   p-value     df     fmi       Inf      Sup
omega   0.07101 0.03153 1.296e-02 135.5 0.4102 0.008657 0.13335
alpha1  0.06513 0.01564 2.215e-05 228.3 0.3084 0.034313 0.09595
beta1   0.92277 0.02262 2.946e-80 141.0 0.4014 0.878051 0.96749
eta11   0.51707 0.14722 2.765e-04 193.5 0.3381 0.226706 0.80744
shape   1.27633 0.06081 3.151e-43 125.8 0.4266 1.155993 1.39667
```

Antes de analizar estos resultados, calculamos el mismo modelo en dos escenarios más: uno el caso de completitud de datos y otro omitiendo los valores NA.

```
# Coeficientes estimados en caso de completitud de datos
        Estimate  Std. Error  t value  Pr(>|t|)
omega    0.07775     0.02398    3.243 1.184e-03
alpha1   0.07440     0.01327    5.607 2.055e-08
beta1    0.91317     0.01748   52.250 0.000e+00
eta11    0.52383     0.10835    4.835 1.334e-06
shape    1.29749     0.04728   27.441 0.000e+00

# Coeficientes estimados omitiendo valores NA
        Estimate  Std. Error  t value  Pr(>|t|)
omega    0.10364     0.03046    3.402 6.678e-04
alpha1   0.08843     0.01584    5.584 2.354e-08
beta1    0.89176     0.02122   42.020 0.000e+00
eta11    0.58178     0.11504    5.057 4.251e-07
shape    1.31566     0.05301   24.820 0.000e+00
```

Estos resultados muestran que, la imputación múltiple ofrece resultados muy similares a los obtenidos en el caso de completitud de datos. Tal como se había predicho en el epígrafe *3.3*, separar y_t en dos series permite capturar la asimetría y la leptocurtosis de la serie original. El coeficiente de asimetría τ es *0.51707* frente al original de *0.52383*; el peso de las colas, ψ, ha aumentado ligeramente, *1.27633* frente a *1.29749*, mientras que cuando eliminamos los NA las colas son más delgadas. El resto de coeficientes resultantes de la imputación son muy cercanos a los reales; sin embargo, ponemos de relieve que se ha suavizado ligeramente α y hay un pequeño aumento en β.

En virtud de este ejemplo intuimos que el método de *imputación mediante separación* ofrece una aproximación comparativa satisfactoria con respecto a la muestra completa y, en todo caso, dicha aproximación es preferible al procedimiento "quirúrgico" cual es la omisión de los NA. Surgen, por consiguiente, dos cuestiones:

- ¿Es ampliable esta primera conclusión a otros títulos?; y ¿cabe la extensión a nivel de índices?
- ¿Hay un límite al porcentaje de NA a partir del cual se invalidara el método?

A lo largo del capítulo 4, por vía inductiva, procuraremos dar respuesta a estos interrogantes.

3.4.3. Manual de la librería

Las páginas 92 - 105 a continuación contienen una doble numeración, una perteneciente a la Tesis (en la parte inferior) y otra que se corresponde con la propia de acuerdo con el las guías de estilo de R (parte superior).

Package 'mists'

June 13, 2013

Type Package

Title Multiple Imputation Simulation for financial time Series

Version 1.0

Date 2012-12-22

Author Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

Maintainer Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

Depends coda,mice,moments,rugarch,timeSeries,xtable,LaplacesDemon

Description This package performs multiple imputation simulation for time series.

License GPL (>= 2)

URL http://canoberlanga.tumblr.com

R topics documented:

93

`mists-package` *Multiple Imputation Simulation for time Series*

Description

Multiple Imputation Simulation for time Series

Details

Package:	mists
Type:	Package
Version:	1.0
Date:	2013-01-15
License:	GPL (>= 2)

Multiple Imputation Simulation for time Series

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com> Maintainer: Sebastian Cano-Berlanga <cano.berlanga@gmail.co

Examples

```
#data(Ibm.r)
#Ibm.mis<-Make.NA(Ibm.r)
#Ibm.imp<-Impute(Ibm.mis$Datum)
#Ibm.imp<-Sample.mids(Ibm.imp)
#analysis<-Garch.mids(Ibm.imp,Zakoian())
#pool<-Rubin.global(analysis,length(Ibm.r))

##Example 2

#data(Ibm.r)
#Ibm.mis<-extreme.values.NA(Ibm.r,4)
#Ibm.imp<-Impute(Ibm.mis)
#Ibm.imp<-Sample.mids(Ibm.imp)
#analysis<-Garch.mids(Ibm.imp,Zakoian())
#pool<-Rubin.global(analysis,length(Ibm.r))
```

`Bayesian` *Bayesian inference*

Description

Performs the Bayesian inference over the quantity of interest.

Usage

```
Bayesian(x)
```

Arguments

x	Output object after using the function `Garch.mids`.

Details

It returns: Mean, Standard Deviation, Percentiles(2.5;50;97.5) plus the HPD Interval.

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

`csGarch.mids` *csGarch Model for Multiply Imputed Data*

Description

Applies the choosen the garch specification to a multiply imputed data set

Usage

```
csGarch.mids(x,y)
```

Arguments

x	Multiply imputed dataset.
y	Garch specification.

Details

The specification must be set following the `rugarch` package. Default solver is `gosolnp`.

Value

It creates a list which stores the vector `Q`, `U` to calculate Rubin's inference. It also stores the matrix `bayesQ` usedto perform bayesian inference on the quantity of interest. Garch model is calculated with `ugarchfit` function of `rugarch` package.

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

References

Engle, R.F. and G.G.J. Lee (1999), A Permanent and Transitory Component Model of Stock Return Volatility, in R.F. Engle and H. White (eds.), Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W.J. Granger, 475-497. Oxford, UK: Oxford University Press.

`Describe` *Descriptive Statistics*

Description

Summary statistics

Usage

```
Describe(x)
```

Arguments

`x`	A data frame, matrix or vector

Value

Returns a data frame with:

Mean
Standard Deviation
Skewness
Kurtosis
Minimum
Percentile 25
Percentile 50
Percentile 75
Maximum

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

Examples

```
data(Ibm.r)
Describe(Ibm.r)
```

`extreme.values` *Extreme Values*

Description

Displays extreme values of time series.

Usage

```
extreme.values(Data, threshold)
```

Arguments

`Data`	Univariate time series
`threshold`	extreme value border

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

`extreme.values.NA` *extreme.values.NA*

Description

Replaces extreme values of a time series by NA

Usage

```
extreme.values.NA(Data, threshold)
```

Arguments

`Data`	Univariate time series
`threshold`	extreme value border

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

`Filter` *Filter*

Description

Filter

Usage

```
Filter(x)
```

Arguments

`x`	Filter

Details

Filter

`Garch.mids` *Garch Model for Multiply Imputed Data*

Description

Applies the choosen the garch specification to a multiply imputed data set

Usage

```
Garch.mids(x,y)
```

Arguments

`x`	Multiply imputed dataset.
`y`	Garch specification.

Details

The specification must be set following the `rugarch` package. Default solver is `gosolnp`

Value

It creates a list which stores the vector `Q`, `U` to calculate Rubin's inference. It also stores the matrix `bayesQ` usedto perform bayesian inference on the quantity of interest. Garch model is calculated with `ugarchfit` function of `rugarch` package.

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

References

Bollerslev, T. (1986), Generalized Autoregressive Conditional Heteroskedasticity, Journal of Econometrics, 31, 307-327.

Nelson, D.B. (1991), Conditional Heteroskedasticity in Asset Returns: A New Approach, Econometrica, 59, 347-370.

Zakoian, J.-M. (1994), Threshold Heteroskedastic Models, Journal of Economic Dynamics and Control, 18, 931-955.

`Hentschel` *Hentschel*

Description

Hentschel

Usage

```
Hentschel()
```

Arguments

`x`	Hentschel

Details

Hentschel

`Ibm.r` *Returns of IBM*

Description

Daily log-returns of IBM from 1998-01-05 to 2013-01-17.

Usage

```
data(Ibm.r)
```

Source

Retrieved from Yahoo Finance with `get.hist.quote` of the `tseries` package.

`Impute` *Multiple Imputation of time series*

Description

This function performs multiple imputation of time series following Cano-Berlanga (2013).

Usage

```
Impute(Data, lags = 5, imp = 5)
```

Arguments

`Data`	Univariate time series with missing obervations
`lags`	Number of lags
`imp`	Number of imputations

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

`Lee.Engle` *csGARCH of Lee and Engle (1999)*

Description

Component Garch specification with GED distribution for the error term.

Usage

```
Lee.Engle()
```

Details

It can be used directly with `garch.mids` and `ugarchfit`.

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

References

G.J. Lee and R.F. Engle. A permanent and transitory component model of stock return volatility. In Cointegration Causality and Forecasting A Festschrift in Honor of Clive WJ Granger, pages 475-497. Oxford University Press, 1999.

Examples

```
ugarchfit(rt(5000,5),spec=Lee.Engle())
```

`Make.NA` *NA creation*

Description

This function introduces artificial random missing observations

Usage

```
Make.NA(Data, p =0.1)
```

Arguments

`Data`	Univariate time series
`p`	Missing ratio, 0.1 by default

Details

It uses a uniform distribution to select the position of the artificial missing.

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

`Nelson` *Egarch of Nelson (1991)*

Description

Exponential Garch(1,1) specification with GED distribution for the error term.

Usage

```
Nelson()
```

Details

It can be used directly with `garch.mids` and `ugarchfit`.

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

References

Nelson, Daniel B. (1991), "Conditional Heteroskedasticity in Asset Returns: A New Approach," Econometrica, 59, 347-370.

Examples

```
ugarchfit(rt(5000,5),spec=Nelson())
```

`Quantile.NA` *Quantile.NA*

Description

Replaces values between two quantiles by NA

Usage

```
Quantile.NA(Data, a, b)
```

Arguments

`Data`	Univariate time series
`a`	Lower quantile
`b`	Upper quantile

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

`Rubin.global` *Rubin's Inference*

Description

This function pools the multiple imputed datasets using inference rules defined by Rubin (1987).

Usage

```
Rubin.global(x, j)
```

Arguments

`x`	Output of `Garch.mids`.
`j`	Degrees of freedom of the original sample.

Details

It uses the definition of Barnard and Rubin (1999) to calculate the degrees of freedom.

Value

The output is a matrix which contains the following information:

`Q`	Average of the quantity of interest
`SE`	Standard Error of the scalar of interest
`t`	t statistic
`p-value`	p-value rejecting the null hypothesis of Q=0
`r`	Increase in the variance due to the presence of NAs
	Degrees of freedom
`lambda`	Fraction of missing information
`Lower`	Lower bound of confidence interval (p=0.95)
`Higher`	Upper bound of confidence interval (p=0.95)

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

References

Barnard, J. and Rubin, D.B. (1999). Small-sample Degrees of Freedom with Multiple Imputation. Biometrika, 86, 948-955.

Rubin, D.B. (1987) Multiple imputation for non-response in surveys. New York: John Wiley & Sons.

`Rubin.scalar` *Rubin's Inference internal command*

Description

Internal command to perform Rubin's inference

Usage

```
Rubin.scalar(Q, U, j)
```

Arguments

`Q`	Individual values of the quantity of interest
`U`	Variance of the quantity of interest
`j`	Degrees of freedom

Details

It uses the definition of Barnard and Rubin (1999) to calculate the degrees of freedom.

Value

The output is a matrix which contains the following information:

Q	Average of the quantity of interest
SE	Standard Error of the scalar of interest
t	t statistic
p-value	p-value rejecting the null hypothesis of Q=0
r	Increase in the variance due to the presence of NAs
	Degrees of freedom
lambda	Fraction of missing information
Lower	Lower bound of confidence interval (p=0.95)
Higher	Upper bound of confidence interval (p=0.95)

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

References

Barnard, J. and Rubin, D.B. (1999). Small-sample Degrees of Freedom with Multiple Imputation. Biometrika, 86, 948-955.

Rubin, D.B. (1987) Multiple imputation for non-response in surveys. New York: John Wiley & Sons.

Sample.mids *Sample construction*

Description

Forms the Multiply Imputed data frame

Usage

```
Sample.mids(x)
```

Arguments

x	Output object of Impute

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

`Zakoian` *Tgarch of Zakoian (1994)*

Description

Threshold Garch(1,1) specification with GED distribution for the error term.

Usage

```
Nelson()
```

Details

It can be used directly with `garch.mids` and `ugarchfit`.

Author(s)

Sebastian Cano-Berlanga <cano.berlanga@gmail.com>

References

Zakoian, J.M. (1994) Threshold heteroskedastic models. Journal of Economic Dynamics and Control 18, 931-955.

Examples

```
ugarchfit(rt(5000,5),spec=Zakoian())
```

Index

3.4.4. Código de las instrucciones principales

En este apartado se expone el código de las instrucciones principales de **mists** (utilizadas para el desarrollo del capítulo *4*). La librería contiene más instrucciones de las que se utilizarán en este trabajo, hecho que facilita la extensión empírica del método descrito en el epígrafe *3.3*.

```
Filter ←
function(x){
    require(rugarch)
  garch=ugarchspec(mean.model = list(armaOrder = c(0,
        0), include.mean = FALSE, external.regressors = NULL,
        archm = FALSE, archpow = 1, arfima = FALSE), variance.model = list(
            garchOrder = c(1,
        1), model = "fGARCH", submodel = "AVGARCH"), distribution.model = "
            norm")
  GARCH←ugarchfit(x,spec=garch)
  S←sigma(GARCH)
  E←x/S
  M=cbind(S,E)
  colnames(M)←c("St","Et")
  output←list(Et=E,St=S,M=M)
  return(output)
  }
}
```

Código 3.1: Instrucción *Filter()*

```
Make.NA ←
function(Data,p=0.1){
T←dim(Data)[1]
round(runif(p*T,1,T))→mis
Data[c(mis),]←NA
output←list(Datum=Data,location=mis)
return(output)
}
```

Código 3.2: Instrucción *Make.NA()*

```
Impute ←
function(Data,lags=5,im=5){
```

```
X<-Data[,1]
E<-Data[,2]

imp <- ABB(X, K=1)
X.imp <- X
X.imp[which(is.na(X.imp))] <- unlist(imp)
St<-as.matrix(X.imp)

embed(St,lags)->S
embed(E,lags)->E

Dim<-dim(E)
a.out <- mice(E,m=im)
IMP<-array(NA,c(Dim,im))

for (i in 1:im){
  IMP[,,i]<-as.matrix(S*complete(a.out,i))
}

output<-list(Imp=IMP,m=im)
return(output)
}
```

Código 3.3: Instrucción *Impute()*

```
Sample.mids <-
function(x){
  m<-x$m
  Sample<-NULL
  for (i in 1:m){
    Sample<-cbind(Sample,x$Imp[,,i])
      }
return(as.matrix(Sample))
}
```

Código 3.4: Instrucción *Sample.mids()*

```
Garch.mids<-
function (x, y=Zakoian(),z="solnp")
{
    m <- dim(x)[2]
    outputQ <- NULL
    outputU <- NULL
    bayesQ <- NULL
    for (i in 1:m) {
      X<-x[,i]
        GARCH <- ugarchfit(X, spec = y,solver=z)
        GARCH.q <- GARCH@fit$coef
        GARCH.u <- GARCH@fit$se.coef
        bayesQ <- rbind(bayesQ, t(as.matrix(c(GARCH@fit$coef,
            uncvariance(GARCH)^0.5, persistence(GARCH), halflife(GARCH)))))
        outputQ <- rbind(outputQ, GARCH.q)
        outputU <- rbind(outputU, GARCH.u)
    }
bayesQ<-as.matrix(bayesQ)
colnames(bayesQ)<-c(rownames(GARCH@fit$matcoef),"Sy","Per","h2l")
ME<-apply(bayesQ,2,mean)
return<-list(Q=outputQ,U=outputU^2,bayesQ=bayesQ,names=rownames(GARCH@fit$
    matcoef),ME=ME)
}
```

Código 3.5: Instrucción *Garch.mids()*

```
Rubin.scalar <-
function(Q,U,j){

    m<-length(Q)
    qbar <- mean(Q)
    ubar <- mean(U)
    b <- var(Q)
    t <- ubar + (m + 1) * b/m
    se<- sqrt(t)
    r <- (1 + 1/m) * b/ubar
    gamma<-(1+1/m)*b/t
    df.complete<-j
    df.large<-(m-1)*(1+1/r)^2


    ### Adjustement Barnard & Rubin (1999)
    df.obs<-df.complete*(df.complete+1)*(1-gamma)/(df.complete+3)
    df<-(1/df.large + 1/df.obs)^(-1)
    ### end adjustment  Barnard & Rubin (1999)
```

```
    f ← (r + 2/(df + 3))/(r + 1)
    res←matrix(NA,nrow=1,ncol=9)
    colnames(res)←c("Q","SE","t","p-value","r","df","fmi","Inf","Sup")
    res[1,1]←qbar
    res[1,2]←se
    res[1,3]←qbar/se
    res[1,4]←pt(abs(qbar/se) ,df,lower.tail=FALSE)
    res[1,5]←r
    res[1,6]←abs(df)
    res[1,7]←f
    res[1,8]←qbar-qt(0.025,df,lower.tail=FALSE)*se
    res[1,9]←qbar+qt(0.025,df,lower.tail=FALSE)*se
    return(res)
}
```

Código 3.6: Instrucción *Rubin.scalar()*

```
Rubin.global ←
function(x,j){
  dim(x[[1]])[2]→k
    sum(dim(x))→l
  full.inference←NULL
  for (i in 1:k){
    inference←Rubin.scalar(x$Q[,i],x$U[,i],j-l)
    full.inference←rbind(full.inference,inference)
    }
  row.names(full.inference)←x$names
  return(full.inference)

}
```

Código 3.7: Instrucción *Rubin.global()*

Capítulo 4

Validación empírica del método propuesto

"No aplicado, de nada sirve el saber"
Calderón de la Barca, Eco y Narciso

En el presente capítulo, tal y como se adelantó en el capítulo *3*, evaluaremos empíricamente el proceso de imputación mediante separación. Para ello proponemos una muestra de firmas cotizadas e índices bursátiles, apartado *4.1*, sobre la que estimaremos los modelos descritos en el epígrafe *4.2*. Las estimaciones obtenidas, epígrafe *4.3*, se utilizarán como *benchmark* de comparación.

Para evaluar el método de imputación mediante separación eliminaremos, **de forma aleatoria**, porcentajes del 10 %, 15 % y 30 % en cada serie temporal de la muestra. Tras este paso, aplicaremos el método propuesto con un número de imputaciones $m = 25$, y seguiremos el mismo esquema utilizado en el apartado *3.4.2*. Los resultados de las imputaciones se muestran en los cuadros del Anexo I.

A efectos de comparar los coeficientes reales y los obtenidos, Anexo II, mediante el método propuesto calcularemos el ratio entre el valor real del coeficiente y valor obtenido por la imputación. En dichos cuadros, el número se resalta en **azul**, cuando los contrastes de significación del coeficiente no coinciden. Por último, en al Anexo III, compararemos gráficamente las densidades de las distribuciones reales de los rendimientos y la volatilidad con las obtenidas tras aplicar la imputación mediante separación.

4.1. Selección de las muestras

A efectos de evaluación del método de imputación mediante separación hemos seleccionado una muestra de ocho firmas cotizadas y siete índices bursátiles (ver cuadros

4.1 y 4.2), para el periodo comprendido entre 2003-01-01 y 2012-12-31, con número de observaciones muy similar.[1]

Con respecto a la muestra, hemos de indicar que no es aleatoria, debido a que no hemos querido únicamente situarnos en un mercado o únicamente en un sector, y dentro de él extraer la muestra aleatoria. Por otra parte, si consideráramos la población completa de índices bursátiles o la población completa de activos que cotizan en las distintas Bolsas mundiales nos hallaríamos ante problemas de eficiencia muy dispar. Por consiguiente, aún siendo conscientes, de un cierto grado de subjetividad hemos realizado la selección utilizando tres criterios:

1. diversidad geográfica
2. diversidad sectorial (en el caso de firmas cotizadas)
3. eficiencia del mercado donde cotizan los activos indicados considerada desde un punto de vista financiero general, es decir, aceptada comúnmente por los expertos

En lo que se refiere al tamaño de las muestras, no tratándose de muestras aleatorias, el número elegido creemos que, en principio, nos permite aceptar de un modo orientativo los resultados del análisis, sobre todo porque tanto las firmas como los índices cotizados conllevan una importancia notoria. Las muestras elegidas fueron las primeras que nos propusimos, y no se han modificado en virtud del resultado de la validación favorable o no.

Ticker	Empresa	Nº Obs
IBM	International Business Machines (EUA)	2516
AAPL	Apple Inc. (EUA)	2516
GE	General Electric (EUA)	2516
TEF	Telefónica S.A. (España)	2602
SAN	Banco Santander (España)	2604
NOVN	Novartis (Suiza)	2549
BMW	Bayerische Motoren Werke AG (Alemania)	2580
EAD	European Aero. Def. and Space (Francia)	2573

Cuadro 4.1: Selección de la muestra de firmas cotizadas

[1] Las series se han descargado directamente en **R** mediante `get.hist.quote()`.

Índice	Nombre	Nº Obs
NASDAQ	NASDAQ Composite (EUA)	2516
SP500	Standard & Poor's 500 Index (EUA)	2516
EUROSTOXX	Dow Jones EURO STOXX 50 (UE)	2557
IBEX	IBEX 35 (España)	2540
SMI	Swiss Market Index (Suiza)	2538
DAX	Xetra DAX (Alemania)	2555
CAC	Cotation Assistée en Continu	2562

CUADRO 4.2: Selección de la muestra de índices de mercado

4.2. Modelos que utilizaremos para las estimaciones

Entre los numerosos modelos GARCH existentes en la literatura hemos elegido: el Exponential GARCH de Nelson [1991], por su importante peso en la literatura empírica y carácter asimétrico; el *asymmetric absolute value* GARCH de Hentschel (1995) por sus dos coeficientes de asimetría; y el *Component* GARCH Lee & Engle (1999), por su descomposición de la varianza condicionada en componentes a corto y largo plazo. La especificación completa dichos modelos, además de la volatilidad condicionada, incluye la media incondicionada (μ) de la serie de rendimientos y el peso de las colas (ψ) de la distribución GED[2] de Nelson [1991] para $y_t|\Omega_{t-1}$.

Exponential GARCH

Nelson [1991] propuso un GARCH en el que se modeliza el logaritmo de σ_t^2. Esta especificación de la varianza condicionada presentaba dos importantes novedades: por un lado no necesita restricción alguna sobre los coeficientes del modelo para garantizar la positividad de σ_t^2, y por otro lado fue el primer modelo que incorporaba un término de asimetría. La dinámica del EGARCH (con $p = q = 1$) es,

$$log\sigma_t^2 = \omega + \beta\, log\, \sigma_{t-1}^2 + \alpha|\varepsilon_{t-1}| + \tau \cdot \varepsilon_{t-1}$$

la persistencia del modelo se define mediante,

$$\phi = \beta$$

[2]Las colas de la función de distribución GED dependen de un parámetro ψ, que mide su peso: cuánto más pequeño es ψ más pesadas son las colas de la distribución y viceversa. Además, esta función de distribución tiene como casos particulares la distribución de Laplace si $\psi = 1$, la distribución Normal si $\psi = 2$ y la distribución uniforme si $\psi \to \infty$.

la varianza incondicionada es,

$$\sigma_y^2 = exp\left\{\frac{\omega}{1-\phi}\right\}$$

y la curva de impacto de las noticias (NIC) se define en función del signo de las innovaciones,

$$NIC = \begin{cases} \sigma_y^{2\beta} exp[\omega] \cdot exp\left[\frac{\tau-\alpha}{\sigma_y}\varepsilon_{t-1}\right] & \text{si } \varepsilon_{t-1} < 0 \\ \sigma_y^{2\beta} exp[\omega] \cdot exp\left[\frac{\tau+\alpha}{\sigma_y}\varepsilon_{t-1}\right] & \text{si } \varepsilon_{t-1} \geqslant 0 \end{cases}$$

Asymmetric absolute value GARCH

Hentschel (1995) amplía el *Threshold* GARCH de Zakoïan (1994) mediante la inclusión de dos coeficientes de asimetría: τ_1, que mide el impacto cuando ε_{t-1} es grande, y τ_2, que mide el impacto cuando ε_{t-1} es pequeño. Esta especificación recibe el nombre de *asymmetric absolute value* GARCH, cuya dinámica (con $p = q = 1$) es,

$$\sigma_t = \omega + \alpha\sigma_{t-1}(|\varepsilon_{t-1} - \tau_2| - \tau_1(\varepsilon_{t-1} - \tau_2)) + \beta\sigma_{t-1}$$

donde $\omega > 0$, $\alpha \geqslant 0$, $\beta \geqslant 0$, $\tau_1 \in [-1, 1]$, y τ_2 no está acotada.

La persistencia del modelo se define mediante,

$$\phi = \alpha \cdot E[|\varepsilon_{t-1} - \tau_2| - \tau_1(\varepsilon_{t-1} - \tau_2)] + \beta$$

la varianza incondicionada es,

$$\sigma_y^2 = \frac{\omega}{1-\phi}$$

y la curva NIC se define en función del signo de las innovaciones,

$$NIC = \begin{cases} \omega + \beta\alpha\sigma_y(|\varepsilon_{t-1} - \tau_2| - \tau_1(\varepsilon_{t-1} - \tau_2)) & \text{si } \varepsilon_{t-1} < 0 \\ \omega + \beta\alpha\sigma_y(|\varepsilon_{t-1} - \tau_2|) & \text{si } \varepsilon_{t-1} \geqslant 0 \end{cases}$$

Component GARCH

Lee & Engle (1999) proponen una especificación de la varianza condicionada dividida en componentes permanente y transitoria. La dinámica propuesta se denomina *component* GARCH e incluye coeficientes que describen los movimientos de la volatilidad que afectan a largo y corto plazo. La ecuación general (con $p = q = 1$) es,

$$\begin{cases} \sigma_t^2 = q_t + \alpha\left(\varepsilon_{t-1}^2 - q_{t-1}\right) + \beta\left(\sigma_{t-1}^2 - q_{t-1}\right) \\ q_t = \omega + \varphi q_{t-1} + \kappa\left(\varepsilon_{t-1}^2 - \sigma_{t-1}^2\right) \end{cases}$$

donde $\omega > 0$, $\alpha \geqslant 0$, $\beta \geqslant 0$, $\varphi \geqslant 0$ y $\kappa \geqslant 0$. En este modelo, q_t, intersección de la

ecuación de la varianza condicionada, es la componente a largo plazo (permanente), mientras que la diferencia $\sigma^2_{t-j} - q_{t-j}$ es la componente a corto plazo (transitoria). Este sistema de ecuaciones puede escribirse en una sola,

$$\sigma^2_t = \alpha\left(\varepsilon^2_{t-1} - q_{t-1}\right) + \underbrace{\beta\left(\sigma^2_{t-1} - q_{t-1}\right)}_{\text{comp. transistoria}} + \underbrace{\omega + \varphi q_{t-1} + \kappa\left(\varepsilon^2_{t-1} - \sigma^2_{t-1}\right)}_{\text{comp. permanente}}$$

Lee & Engle (1999) deducen cuáles son las condiciones necesarias para garantizar la positividad de σ^2_t y definen la persistencia, ϕ, para las componentes transitoria y permanente,

$$\phi = \begin{cases} \alpha + \beta & \text{para la componente } \textit{transitoria} \\ \varphi & \text{para la componente } \textit{permanente} \end{cases}$$

Para obtener la varianza incondicionada utilizamos la demostración propuesta en Ghalanos (2013),

$$\begin{aligned} E_{t-1}\left[q_{t+n}\right] &= \omega + \varphi E_{t-1}\left[q_{t+n-1}\right] + \kappa E_{t-1}\left[\varepsilon^2_{t+n-j} - \sigma^2_{t+n-j}\right] \\ &= \omega + \varphi E_{t-1}\left[q_{t+n-1}\right] \\ &= \omega + \varphi\left[\omega + \varphi E_{t-1}\left[q_{t+n-2}\right]\right] \\ &= \dots \\ &= \left(1 + p + \dots + \varphi^{n-1}\right)\omega + \varphi^n q_t \\ &= \frac{1-\varphi^n}{1-\varphi}\omega + \varphi^n q_t \end{aligned}$$

Así, cuando $n \to \infty$, obtenemos la varianza incondicionada,

$$E_{t-1}\left[\sigma^2_{t+n}\right] = E_{t-1}\left[q_{t+n}\right] = \sigma^2_y = \frac{\omega}{1-\varphi}$$

A partir de este resultado definimos la NIC del Component GARCH,

$$NIC = \omega + \varphi\sigma^2_y + (\kappa + \alpha)\cdot\left(\varepsilon^2_{t-1} - \sigma^2_y\right)$$

4.3. Comentarios generales sobre los resultados

Tal y como se ha explicado a lo largo del presente capítulo, para evaluar el método de imputación mediante separación, estimamos los modelos tratados en el epígrafe *4.2* sobre las series temporales seleccionadas en la sección *4.1*. La especificación completa de los modelos GARCH estimados es la siguiente:

1. Modelo EGARCH

$$\begin{aligned} y_t &= \mu + \sigma_t \varepsilon_t \\ y_t|\Omega_{t-1} &\sim GED(\mu,\sigma_t,\psi) \\ log\sigma_t^2 &= \omega + \beta\, log\, \sigma_{t-1}^2 + \alpha|\varepsilon_{t-1}| + \tau \cdot \varepsilon_{t-1} \end{aligned}$$

2. Modelo AAVGARCH

$$\begin{aligned} y_t &= \mu + \sigma_t \varepsilon_t \\ y_t|\Omega_{t-1} &\sim GED(\mu,\sigma_t,\psi) \\ \sigma_t &= \omega + \alpha\sigma_{t-1}(|\varepsilon_{t-1} - \tau_2| - \tau_1(\varepsilon_{t-1} - \tau_2)) + \beta\sigma_{t-1} \end{aligned}$$

3. Modelo CGARCH

$$\begin{aligned} y_t &= \mu + \sigma_t \varepsilon_t \\ y_t|\Omega_{t-1} &\sim GED(\mu,\sigma_t,\psi) \\ \sigma_t^2 &= q_t + \alpha\left(\varepsilon_{t-1}^2 - q_{t-1}\right) + \beta\left(\sigma_{t-1}^2 - q_{t-1}\right) \\ q_t &= \omega + \varphi q_{t-1} + \kappa\left(\varepsilon_{t-1}^2 - \sigma_{t-1}^2\right) \end{aligned}$$

dónde μ es la media incondicionada de los rendimientos y ψ es el peso de las colas de la distribución GED. Las estimaciones de los modelos sin valores NA se muestran en los cuadros *4.3 - 4.8* del Anexo I; las obtenidas tras aplicar la imputación mediante separación están disponibles en los cuadros *4.9 - 4.26* del Anexo I. Las comparaciones entre las estimaciones normales y las resultantes de las imputaciones se muestran en los cuadros *4.27 - 4.35* del Anexo II. Por último, para ilustrar gráficamente los resultados obtenidos por la imputación mediante separación hemos incluido los siguientes gráficos en el Anexo III:

figuras 4.1 - 4.15: comparan las densidades de la distribución teórica de los rendimientos con la obtenida cuando la serie temporal presenta un 15 % de valores NA.

figuras 4.16 - 4.25: comparan las densidades reales de σ_t con las obtenidas por el algoritmo ABB cuando la serie temporal presenta un 15 % de valores NA.

A continuación presentamos los resultados que se desprenden a partir de los cuadros del Anexo I y II:

EGARCH para firmas cotizadas

La media incondicionada se distorsiona a medida que aumenta el porcentaje de valores NA, especialmente en las empresas GE, SAN, IBM y BMW. Respecto a ω observamos que se mantiene por debajo de valor real, salvo EAD (ver cuadro *4.27*) y NOVARTIS; éste último tiene una magnitud superior en todos los escenarios (ver cuadro *4.27 - 4.29*). Los parámetros relativos al impacto de las noticias , α y τ, se mantienen por debajo de su magnitud real, comportamiento que se acentúa cuando aumenta el porcentaje de valores NA. El coeficiente β exhibe gran precisión, ya que su la magnitud tras las imputaciones se sitúa alrededor de los valores reales. El peso de las colas, ψ, y la volatilidad incondicionada, σ_y, disfrutan de muy poco error; aunque éste va aumentando a medida que falta más información.

EGARCH para índices de mercado

La media incondicionada no presenta un patrón específico de error en los escenarios del 10 % y 15 % de valores NA, en cambio, μ es siempre más pequeña en todos los casos cuando el porcentaje de NA es del 30 %. La constante ω exhibe valores menores que los reales (salvo en DAX, ver cuadros *4.27 - 4.29*), hecho que se acentúa a medida que eliminamos información. El coeficiente α se mantiene en todos los casos por debajo del valor real; lo mismo sucede con τ, excepto cuando los valores NA alcanzan el 30 %, donde sucede lo contrario (ver cuadro *4.29*). El coeficiente β vuelve a ser muy fiel al original, alcanzando magnitudes próximas a las reales. El parámetro ψ exhibe unos resultados similares al de β, aunque el error aumenta si hay mayor presencia de valores NA.

AAVGARCH para firmas cotizadas

El coeficiente μ no tiene un patrón común de error para los escenarios contemplados, sin embargo destacamos la reducción de su valor para GE en todos los casos (ver cuadros *4.15 - 4.17, 4.30 - 4.32*), y el fallo en el contraste de significación para Novartis y BMW (cuadro *4.32*). La estimación de ω se mantiene por debajo de los valores reales, salvo para Novartis (ver cuadro *4.31*). La magnitud de α también es inferior a la estimada cuando los datos no contienen valores NA, además nos gustaría poner de manifiesto el elevado error que este coeficiente presenta un todas las situaciones. Los coeficientes relativos a la asimetría, τ_1 y τ_2, son insatisfactorios: no exhiben un comportamiento específico de de error, presentan cambios de signo en los valores estimados y los contrastes de significación fallan en varias ocasiones (especialmente en τ_2). Los coeficientes β, ψ y σ_y muestran un comportamiento parecido: son muy exactos y su magnitud es mayor que la original (exceptuando algún caso).

AAVGARCH para índices de mercado

La magnitud de la media incondicionada imputada es superior a sus valores reales (exceptuando DAX), y a medida que aumenta el porcentaje de valores NA los test de significación se alteran (IBEX, SMI y CAC). Respecto a ω podemos afirmar que las magnitudes del error son elevadas (enfatizamos el caso de Apple, ver cuadros *4.31 - 4.32*), no existe un patrón definido de error y los test de significación son correctos. El comportamiento de α es mejor que en los escenarios comentados hasta aquí; no existe regularidades en los errores y las diferencias respecto a los valores originales no son tan elevadas como las expuestas con anterioridad. Los buenos resultados de α contrastan con los valores de τ_1 y τ_2, que exhiben el mismo comportamiento que las firmas cotizadas para este modelo. El peso de las colas, ψ, es de nuevo mayor que el estimado cuando los datos están completos; su error no es elevado pero aumenta a medida que hay más NA.

CGARCH para firmas cotizadas

La media incondicionada alcanza niveles inferiores a los reales en la mayoría de escenarios y, el test de significación falla en TEF y SAN. La constante ω se mantiene por debajo de los valores reales cuando el porcentaje de NA es del 10 %; no obstante, al aumentar el porcentaje de valores NA el error no exhibe el mismo comportamiento. El coeficiente α no muestra un patrón claro de error, salvo cuando los valores NA son el 30 % de la muestra: en este caso α se mantiene por debajo de los valores originales, aumenta el error y el test de significación falla en varias ocasiones. Las estimaciones de β son muy aceptables, aunque señalamos el fallo en los test de significación para BMW, GE y EAD (ver cuadro *4.35*). En línea con este resultado está el coeficiente φ, de elevadísima precisión en todos los escenarios. Por último, tal y como ha venido sucediendo hasta aquí, las colas de la distribución condicionada son más pesadas que las originales.

CGARCH para índices de mercado

La media incondicionada tiene resultados aceptables, únicamente falla el test de significación en el índice CAC (ver cuadro *4.35*). La constante ω no exhibe un comportamiento definido respecto al error y los test de significación fallan varia veces. Los resultados respecto a α muestran que sus estimaciones tienen magnitudes menores que las originales (resaltamos la notable diferencia de valores en SP500, ver cuadros *4.33 - 4.35*). El coeficiente β exhibe un comportamiento similar al de las firmas cotizadas, pero el contraste de significación falla en SP500 (cuadros *4.34 - 4.35*) y CAC (cuadro *4.35*). La estimación de ϕ es, de nuevo, muy precisa, prácticamente idéntica a la original. Por último, nuevamente las colas de la distribución condicionada son más pesadas que en el caso de tener datos completos (el peso aumenta a medida que se eleva el porcentaje de

valores NA).

Resultados globales

- La media incondicionada, μ, no sigue un patrón específico de error. Además es el coeficiente para el que, en términos globales, más ocasiones falla el test de significación. En términos del modelo GARCH, éste es un hecho esperable (pero no necesariamente satisfactorio), pues μ depende de los valores de los valores extremos de y_t.

- La constante ω se mantiene, salvo excepciones, por debajo de sus valores reales. Los test de significación funcionan adecuadamente en la mayoría de ocasiones. También destacamos que el patrón obtenido es independiente del nivel de valores NA presente en y_t.

- Los coeficientes relativos al impacto de las innovaciones muestran frecuentemente una reducción respecto al valor real. La disminución de estos parámetros aumenta a medida que incrementamos el porcentaje de NA.

- El valor de β es muy preciso, acercándose mucho al valor original, aunque las estimaciones imputadas ofrecen una magnitud mayor respecto al valor real. De hecho, el error en este coeficiente es del uno o dos por ciento en la mayoría de casos. En cuanto a la incertidumbre, λ aumenta a medida que se eleva el porcentaje de valores NA en todos los modelos y series.

- El parámetro β es, en el EGARCH y AAVGARCH, uno de los más satisfactorios. Sus estimaciones puntuales son muy precisas, pero su magnitud es mayor en todos los escenarios. Este resultado es extensible al modelo de Lee & Engle [1999], donde φ muestra el mismo comportamiento que β. Por consiguiente, podemos concluir que las series imputadas aumentan la tasa a la que decaen los impactos en la varianza condicionada.

- Los parámetros que capturan la asimetría son satisfactorios en el modelo EGARCH, pero su magnitud se reduce. En el modelo AAVGARCH los resultados son mixtos: por un lado τ_1 es satisfactorio aunque esporádicamente el test de significación es incoherente con el valor real (los contrastes sobre τ_1 fallan menos en índices bursátiles), y por otro lado el coeficiente τ_2 revela un desajuste importante con los resultados reales. Este resultado es coherente con las magnitudes de λ, muy elevadas en todos los casos.

- El peso de las colas de la distribución condicionada es muy fiel al estimado con los datasets completos. El parámetro ψ tiene un error muy pequeño, pero su valor es siempre algo menor comparado con el real.

- La persistencia de los modelos, ϕ, y la volatilidad incondicionada, σ_y, poseen una elevada precisión con respecto a las estimaciones originales. La tendencia general que siguen estos parámetros reflejan un aumento de ϕ con su correspondiente efecto sobre σ_y.

Anexo I: Cuadros de resultados

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
IBM	0.0335	0.0086	-0.0639	0.9760	0.1577	1.3103	1.1954	0.9760	28.4788
	0.1187	0.0507	0.0000	0.0000	0.0000	0.0000			
APPLE	0.1779	0.0462	-0.0612	0.9709	0.1464	1.3000	2.2130	0.9709	23.4924
	0.0000	0.0019	0.0001	0.0000	0.0000	0.0000			
GE	0.0029	0.0051	-0.0517	0.9932	0.1215	1.3593	1.4609	0.9932	102.1261
	0.8859	0.0641	0.0000	0.0000	0.0000	0.0000			
TEF	0.0150	0.0126	-0.0846	0.9735	0.1823	1.3995	1.2686	0.9735	25.8251
	0.4765	0.0105	0.0000	0.0000	0.0000	0.0000			
SAN	0.0145	0.0179	-0.1233	0.9826	0.1703	1.3892	1.6739	0.9826	39.5709
	0.5198	0.0024	0.0000	0.0000	0.0000	0.0000			
NOVARTIS	0.0336	0.0033	-0.0541	0.9871	0.0962	1.3499	1.1367	0.9871	53.5128
	0.1161	0.1875	0.0000	0.0000	0.0000	0.0000			
BMW	0.0436	0.0123	-0.0364	0.9891	0.1443	1.4285	1.7615	0.9891	63.3368
	0.1418	0.0139	0.0029	0.0000	0.0000	0.0000			
EAD	0.0532	0.0218	-0.0585	0.9854	0.1273	1.4102	2.1108	0.9854	47.2296
	0.1361	0.0048	0.0000	0.0000	0.0000	0.0000			

CUADRO 4.3: Estimaciones del modelo de Nelson [1991] para firmas cotizadas. La primera fila de cada índice muestra los valores de los coeficientes, mientras que la segunda son los *p-values*.

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0552	0.0025	-0.0986	0.9842	0.1131	1.5152	1.0832	0.9842	43.5408
	0.0066	0.4055	0.0000	0.0000	0.0000	0.0000			
SP500	0.0495	-0.0059	-0.1261	0.9867	0.1106	1.3740	0.8025	0.9867	51.7794
	0.0017	0.0462	0.0000	0.0000	0.0000	0.0000			
EUROSTOXX	0.0125	0.0029	-0.1528	0.9826	0.1047	1.6054	1.0868	0.9826	39.5001
	0.5150	0.4423	0.0000	0.0000	0.0000	0.0000			
IBEX	0.0525	0.0015	-0.1151	0.9842	0.1230	1.5043	1.0498	0.9842	43.5080
	0.0063	0.6406	0.0000	0.0000	0.0000	0.0000			
SMI	0.0286	-0.0058	-0.1361	0.9802	0.1156	1.6243	0.8644	0.9802	34.6736
	0.0697	0.0645	0.0000	0.0000	0.0000	0.0000			
DAX	0.0667	0.0015	-0.1257	0.9837	0.1203	1.4729	1.0467	0.9837	42.2095
	0.0007	0.6705	0.0000	0.0000	0.0000	0.0000			
CAC	0.0177	0.0034	-0.1556	0.9810	0.1045	1.6363	1.0936	0.9810	36.1793
	0.3622	0.3794	0.0000	0.0000	0.0000	0.0000			

CUADRO 4.4: Estimaciones del modelo de Nelson [1991] para índices de mercado. La primera fila de cada índice muestra los valores de los coeficientes, mientras que la segunda son los *p-values*.

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
IBM	0.0293	0.0396	0.0976	0.8929	0.3671	0.0574	1.3200	1.2510	0.9683	21.5491
	0.1659	0.0001	0.0000	0.0000	0.0015	0.4651	0.0000			
APPLE	0.1714	0.0803	0.0900	0.8819	0.0541	0.5470	1.3029	2.2847	0.9649	19.3761
	0.0000	0.0008	0.0000	0.0000	0.5897	0.0000	0.0000			
GE	0.0001	0.0128	0.0721	0.9366	0.4210	0.0345	1.3628	1.6051	0.9921	86.8936
	0.9725	0.0053	0.0000	0.0000	0.0000	0.4259	0.0000			
TEF	0.0034	0.0359	0.1257	0.8613	-0.0333	0.6234	1.3854	1.5004	0.9761	28.6385
	0.8380	0.0003	0.0000	0.0000	0.6542	0.0000	0.0000			
SAN	0.0172	0.0268	0.0998	0.9287	0.8665	-0.2638	1.3996	1.7892	0.9850	45.8419
	0.4658	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000			
NOVARTIS	0.0354	0.0179	0.0588	0.9577	0.8311	-0.4830	1.3525	1.1930	0.9850	45.7306
	0.0977	0.0078	0.0000	0.0000	0.0000	0.0000	0.0000			
BMW	0.0418	0.0224	0.0784	0.9294	0.3080	-0.0478	1.4269	1.8705	0.9880	57.5716
	0.1438	0.0022	0.0000	0.0000	0.0017	0.4682	0.0000			
EAD	0.0520	0.0343	0.0720	0.9257	0.3545	0.1454	1.3934	2.2505	0.9848	45.1220
	0.1552	0.0024	0.0000	0.0000	0.0026	0.0005	0.0000			

CUADRO 4.5: Estimaciones del modelo de Hentschel [1995] para firmas cotizadas. La primera fila de cada índice muestra los valores de los coeficientes, mientras que la segunda son los *p-values*.

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0558	0.0233	0.0682	0.9404	1.0000	-0.2200	1.5175	1.1327	0.9794	33.2896
	0.0066	0.0001	0.0000	0.0000	0.0000	0.0001	0.0000			
SP500	0.0457	0.0174	0.0702	0.9290	1.0000	-0.0057	1.3839	0.9499	0.9817	37.5723
	0.0031	0.0000	0.0000	0.0000	0.0000	0.9069	0.0000			
EUROSTOXX	0.0017	0.0239	0.0671	0.9028	1.0000	0.3531	1.6133	1.3492	0.9823	38.7617
	0.9258	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000			
IBEX	0.0471	0.0193	0.0694	0.9363	1.0000	-0.1000	1.5154	1.1408	0.9830	40.5188
	0.0169	0.0000	0.0000	0.0000	0.0000	0.2884	0.0000			
SMI	0.0224	0.0216	0.0699	0.9017	0.8633	0.3227	1.6383	1.0146	0.9788	32.2872
	0.1589	0.0000	0.0000	0.0000	0.0000	0.0004	0.0000			
DAX	0.0573	0.0242	0.0762	0.8930	0.5682	0.4968	1.4801	1.2983	0.9814	36.8593
	0.0033	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000			
CAC	0.0053	0.0260	0.0685	0.8992	1.0000	0.3567	1.6499	1.3534	0.9808	35.7192
	0.7902	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000			

CUADRO 4.6: Estimaciones del modelo de Hentschel [1995] para índices de mercado. La primera fila de cada índice muestra los valores de los coeficientes, mientras que la segunda son los *p-values*.

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
IBM	0.0479	0.0173	0.0596	0.8325	0.9899	0.0363	1.3046	1.3086	0.9899	68.2659
	0.0268	0.0114	0.0021	0.0000	0.0000	0.0015	0.0000			
APPLE	0.1951	0.0310	0.0526	0.7309	0.9945	0.0317	1.2918	2.3823	0.9945	126.3674
	0.0000	0.0702	0.0119	0.0000	0.0000	0.0003	0.0000			
GE	0.0209	0.0082	0.0504	0.7974	0.9978	0.0423	1.3521	1.9234	0.9978	311.3675
	0.3872	0.0259	0.0132	0.0000	0.0000	0.0000	0.0000			
TEF	0.0345	0.0087	0.0826	0.8909	0.9960	0.0137	1.3824	1.4690	0.9960	170.9160
	0.1312	0.3648	0.0000	0.0000	0.0000	0.4500	0.0000			
SAN	0.0607	0.0086	0.0814	0.8607	0.9985	0.0428	1.3368	2.3850	0.9985	457.7579
	0.0296	0.0906	0.0001	0.0000	0.0000	0.0227	0.0000			
NOVARTIS	0.0450	0.0139	0.0736	0.5902	0.9906	0.0300	1.3352	1.2111	0.9906	73.0252
	0.0188	0.0150	0.0035	0.0000	0.0000	0.0001	0.0000			
BMW	0.0618	0.0108	0.0347	0.9362	0.9975	0.0325	1.4163	2.0896	0.9975	279.7552
	0.0385	0.0929	0.0297	0.0000	0.0000	0.0196	0.0000			
EAD	0.0890	0.0512	0.0625	0.6935	0.9918	0.0547	1.3807	2.5048	0.9918	84.6648
	0.0168	0.0348	0.0204	0.0008	0.0000	0.0004	0.0000			

CUADRO 4.7: Estimaciones del modelo de Lee & Engle [1999] para firmas cotizadas. La primera fila muestra los valores de los coeficientes, mientras que la segunda son los *p-values*.

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0856	0.0197	0.0048	0.9835	0.9883	0.0660	1.4804	1.2967	0.9883	58.9204
	0.0000	0.0011	0.9648	0.0000	0.0000	0.5471	0.0000			
SP500	0.0745	0.0128	0.0004	0.9899	0.9903	0.0814	1.3256	1.1490	0.9903	70.9812
	0.0000	0.0008	0.9848	0.0000	0.0000	0.0003	0.0000			
EUROSTOXX	0.0618	0.0051	0.0653	0.9082	0.9975	0.0290	1.4623	1.4196	0.9975	272.8240
	0.0021	0.1390	0.0005	0.0000	0.0000	0.1245	0.0000			
IBEX	0.0869	0.0033	0.0737	0.8922	0.9983	0.0253	1.4220	1.4140	0.9983	413.6550
	0.0000	0.1203	0.0000	0.0000	0.0000	0.0623	0.0000			
SMI	0.0617	0.0062	0.0767	0.8794	0.9944	0.0348	1.5154	1.0509	0.9944	123.5598
	0.0001	0.0474	0.0001	0.0000	0.0000	0.0556	0.0000			
DAX	0.1043	0.0071	0.0459	0.9326	0.9975	0.0423	1.3799	1.6815	0.9975	276.2549
	0.0000	0.1553	0.0549	0.0000	0.0000	0.0994	0.0000			
CAC	0.0681	0.0054	0.0675	0.9023	0.9972	0.0286	1.5082	1.3809	0.9972	246.4602
	0.0006	0.1219	0.0001	0.0000	0.0000	0.1081	0.0000			

CUADRO 4.8: Estimaciones del modelo de Lee & Engle [1999] para índices de mercado. La primera fila muestra los valores de los coeficientes, mientras que la segunda son los *p-values*.

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
IBM	0.0282	0.0049	-0.0457	0.9864	0.1100	1.2837	1.1963	0.9864	50.7768
p-val	0.1065	0.0564	0.0007	0.0000	0.0000	0.0000			
λ	0.1200	0.0693	0.2153	0.1604	0.2293	0.1522			
APPLE	0.1753	0.0316	-0.0418	0.9805	0.1156	1.2688	2.2512	0.9805	35.1607
p-val	0.0000	0.0170	0.0038	0.0000	0.0000	0.0000			
λ	0.1072	0.1953	0.1669	0.2112	0.2140	0.2945			
GE	0.0036	0.0041	-0.0424	0.9947	0.1056	1.2927	1.4687	0.9947	130.1852
p-val	0.4287	0.0534	0.0001	0.0000	0.0000	0.0000			
λ	0.1822	0.0848	0.1746	0.1215	0.2044	0.3078			
TEF	0.0158	0.0128	-0.0691	0.9745	0.1745	1.3624	1.2865	0.9745	26.8873
p-val	0.2593	0.0073	0.0001	0.0000	0.0000	0.0000			
λ	0.2652	0.1045	0.3335	0.1408	0.2701	0.2427			
SAN	0.0268	0.0165	-0.1021	0.9849	0.1569	1.3378	1.7285	0.9849	45.4869
p-val	0.1707	0.0031	0.0000	0.0000	0.0000	0.0000			
λ	0.1809	0.1225	0.2820	0.2538	0.3242	0.4270			
NOVARTIS	0.0321	0.0034	-0.0386	0.9877	0.0995	1.3243	1.1475	0.9877	56.0189
p-val	0.0654	0.1117	0.0014	0.0000	0.0000	0.0000			
λ	0.0846	0.1026	0.2771	0.1838	0.2207	0.1783			
BMW	0.0375	0.0112	-0.0341	0.9903	0.1346	1.4464	1.7815	0.9903	70.9327
p-val	0.1048	0.0113	0.0024	0.0000	0.0000	0.0000			
λ	0.0691	0.0731	0.1636	0.1034	0.1905	0.1164			
EAD	0.0565	0.0225	-0.0466	0.9855	0.1181	1.2906	2.1751	0.9855	47.5087
p-val	0.0748	0.0082	0.0007	0.0000	0.0000	0.0000			
λ	0.0552	0.0913	0.2698	0.0954	0.1318	0.4125			

CUADRO 4.9: Resultados del modelo de Nelson [1991] para firmas cotizadas, porcentaje de NA=10%

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
IBM	0.0348	0.0056	-0.0427	0.9849	0.1231	1.2744	1.2033	0.9849	45.6025
p-val	0.0649	0.0640	0.0049	0.0000	0.0001	0.0000			
λ	0.1434	0.1443	0.3044	0.2452	0.3402	0.3921			
APPLE	0.1385	0.0480	-0.0499	0.9705	0.1403	1.2410	2.2557	0.9705	23.1535
p-val	0.0005	0.0107	0.0025	0.0000	0.0000	0.0000			
λ	0.1410	0.2453	0.1585	0.2398	0.2292	0.4077			
GE	0.0140	0.0034	-0.0492	0.9951	0.0999	1.3447	1.4153	0.9951	140.8270
p-val	0.2816	0.0789	0.0000	0.0000	0.0000	0.0000			
λ	0.1713	0.0605	0.1858	0.1807	0.3166	0.2681			
TEF	0.0264	0.0132	-0.0685	0.9771	0.1630	1.3351	1.3340	0.9771	29.9214
p-val	0.1700	0.0057	0.0001	0.0000	0.0000	0.0000			
λ	0.3129	0.0834	0.3424	0.1383	0.2039	0.4786			
SAN	0.0089	0.0136	-0.0941	0.9880	0.1430	1.2363	1.7717	0.9880	57.4860
p-val	0.3379	0.0066	0.0000	0.0000	0.0000	0.0000			
λ	0.2109	0.1920	0.2656	0.2558	0.4180	0.4469			
NOVARTIS	0.0448	0.0035	-0.0519	0.9870	0.0919	1.2900	1.1448	0.9870	53.0065
p-val	0.0252	0.0949	0.0001	0.0000	0.0000	0.0000			
λ	0.1804	0.0780	0.3200	0.1226	0.1816	0.3531			
BMW	0.0602	0.0102	-0.0321	0.9910	0.1315	1.4418	1.7630	0.9910	76.9510
p-val	0.0318	0.0119	0.0075	0.0000	0.0000	0.0000			
λ	0.1338	0.0421	0.2331	0.0572	0.0988	0.2944			
EAD	0.0740	0.0199	-0.0388	0.9870	0.1093	1.3213	2.1479	0.9870	53.0476
p-val	0.0320	0.0078	0.0019	0.0000	0.0000	0.0000			
λ	0.1544	0.1709	0.2321	0.1860	0.2875	0.3939			

CUADRO 4.10: Resultados del modelo de Nelson [1991] para firmas cotizadas, porcentaje de NA=15%

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
IBM	0.0495	0.0033	-0.0193	0.9908	0.1033	1.2513	1.1979	0.9908	74.9090
p-val	0.0195	0.0958	0.1030	0.0000	0.0000	0.0000			
λ	0.2516	0.1390	0.4003	0.2598	0.2730	0.2971			
APPLE	0.1654	0.0312	-0.0374	0.9810	0.1101	1.2843	2.2724	0.9810	36.0979
p-val	0.0002	0.0279	0.0316	0.0000	0.0000	0.0000			
λ	0.2869	0.3320	0.4757	0.3256	0.2827	0.3175			
GE	-0.0043	0.0029	-0.0345	0.9968	0.0975	1.1979	1.5889	0.9968	213.1911
p-val	0.4342	0.0972	0.0057	0.0000	0.0000	0.0000			
λ	0.3509	0.1302	0.4320	0.2047	0.2404	0.3337			
TEF	0.0162	0.0081	-0.0576	0.9837	0.1311	1.3197	1.2818	0.9837	42.2827
p-val	0.2875	0.0311	0.0001	0.0000	0.0000	0.0000			
λ	0.5627	0.1911	0.2827	0.2791	0.3343	0.3302			
SAN	0.0041	0.0097	-0.0401	0.9919	0.1425	1.1475	1.8114	0.9919	84.8228
p-val	0.4037	0.0227	0.0061	0.0000	0.0000	0.0000			
λ	0.2997	0.1312	0.3541	0.1422	0.2953	0.4182			
NOVARTIS	0.0572	0.0034	-0.0274	0.9885	0.0904	1.2687	1.1572	0.9885	60.0707
p-val	0.0148	0.1195	0.0472	0.0000	0.0000	0.0000			
λ	0.3427	0.1754	0.4648	0.2494	0.3148	0.4608			
BMW	0.0481	0.0079	-0.0128	0.9931	0.1177	1.3429	1.7827	0.9931	100.7627
p-val	0.0860	0.0376	0.1658	0.0000	0.0000	0.0000			
λ	0.2817	0.0838	0.3142	0.1044	0.2133	0.5588			
EAD	0.0847	0.0147	-0.0283	0.9908	0.0934	1.3271	2.2350	0.9908	75.3301
p-val	0.0520	0.0221	0.0437	0.0000	0.0001	0.0000			
λ	0.4392	0.2795	0.5918	0.2744	0.4387	0.7467			

CUADRO 4.11: Resultados del modelo de Nelson [1991] para firmas cotizadas, porcentaje de NA=30%

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0661	0.0019	-0.0730	0.9885	0.1100	1.4452	1.0864	0.9885	60.0818
p-val	0.0012	0.2377	0.0000	0.0000	0.0000	0.0000			
λ	0.1155	0.0590	0.2031	0.2383	0.2504	0.1437			
SP500	0.0494	-0.0041	-0.0943	0.9905	0.1013	1.2612	0.8027	0.9905	72.6741
p-val	0.0011	0.0433	0.0000	0.0000	0.0000	0.0000			
λ	0.1021	0.0382	0.3076	0.2358	0.3784	0.3584			
EUROSTOXX	0.0189	0.0026	-0.1006	0.9876	0.1251	1.4279	1.1078	0.9876	55.7505
p-val	0.1810	0.2116	0.0000	0.0000	0.0000	0.0000			
λ	0.1236	0.0529	0.2712	0.1574	0.2477	0.3103			
IBEX	0.0426	0.0006	-0.0972	0.9883	0.1069	1.4095	1.0254	0.9883	59.0404
p-val	0.0298	0.4156	0.0000	0.0000	0.0000	0.0000			
λ	0.3175	0.0504	0.2788	0.1265	0.2190	0.4339			
SMI	0.0262	-0.0056	-0.1166	0.9796	0.1082	1.5196	0.8716	0.9796	33.6399
p-val	0.0769	0.0340	0.0000	0.0000	0.0000	0.0000			
λ	0.2243	0.1130	0.1945	0.1624	0.2501	0.3053			
DAX	0.0749	0.0021	-0.0940	0.9864	0.1247	1.3708	1.0805	0.9864	50.8037
p-val	0.0004	0.2542	0.0000	0.0000	0.0000	0.0000			
λ	0.2178	0.0367	0.3111	0.1186	0.2534	0.4447			
CAC	0.0224	0.0033	-0.1155	0.9842	0.1200	1.4687	1.1091	0.9842	43.5584
p-val	0.1472	0.1775	0.0000	0.0000	0.0000	0.0000			
λ	0.0903	0.0367	0.2998	0.0950	0.2842	0.3833			

CUADRO 4.12: Resultados del modelo de Nelson [1991] para índices de mercado, porcentaje de NA=10%

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0586	0.0015	-0.0577	0.9904	0.0945	1.3848	1.0795	0.9904	71.6885
p-val	0.0036	0.2554	0.0000	0.0000	0.0000	0.0000			
λ	0.1295	0.0784	0.2639	0.2026	0.3626	0.2396			
SP500	0.0513	-0.0036	-0.0805	0.9921	0.1120	1.2384	0.7955	0.9921	86.8750
p-val	0.0017	0.0644	0.0000	0.0000	0.0000	0.0000			
λ	0.1792	0.0377	0.3093	0.2091	0.2822	0.3998			
EUROSTOXX	0.0287	0.0018	-0.0907	0.9891	0.1056	1.4099	1.0873	0.9891	62.9603
p-val	0.1020	0.2556	0.0000	0.0000	0.0000	0.0000			
λ	0.2436	0.0347	0.3046	0.2115	0.3064	0.4266			
IBEX	0.0407	0.0004	-0.0844	0.9890	0.1021	1.4207	1.0157	0.9890	62.6998
p-val	0.0344	0.4420	0.0000	0.0000	0.0000	0.0000			
λ	0.2603	0.0540	0.4087	0.3689	0.3272	0.4193			
SMI	0.0152	-0.0043	-0.0956	0.9862	0.0965	1.4592	0.8544	0.9862	49.8527
p-val	0.2018	0.0383	0.0000	0.0000	0.0000	0.0000			
λ	0.2273	0.0586	0.3173	0.2050	0.4196	0.2901			
DAX	0.0835	0.0023	-0.0827	0.9868	0.1134	1.3282	1.0895	0.9868	52.3363
p-val	0.0002	0.2262	0.0000	0.0000	0.0000	0.0000			
λ	0.2580	0.0645	0.3852	0.2589	0.3305	0.5514			
CAC	0.0294	0.0031	-0.1075	0.9866	0.1228	1.5240	1.1206	0.9866	51.2371
p-val	0.0797	0.1723	0.0000	0.0000	0.0000	0.0000			
λ	0.1345	0.0302	0.2605	0.1640	0.2687	0.2876			

CUADRO 4.13: Resultados del modelo de Nelson [1991] para índices de mercado, porcentaje de NA=15%

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0618	0.0023	-0.0390	0.9911	0.1111	1.3397	1.1390	0.9911	77.4525
p-val	0.0077	0.1774	0.0058	0.0000	0.0000	0.0000			
λ	0.3236	0.0879	0.4189	0.2593	0.2770	0.3784			
SP500	0.0632	-0.0016	-0.0505	0.9909	0.1352	1.1828	0.9126	0.9909	75.8899
p-val	0.0003	0.2563	0.0020	0.0000	0.0000	0.0000			
λ	0.3106	0.0730	0.3529	0.1438	0.2245	0.3523			
EUROSTOXX	0.0238	0.0021	-0.0630	0.9914	0.1078	1.3207	1.1317	0.9914	80.1876
p-val	0.1763	0.2050	0.0002	0.0000	0.0000	0.0000			
λ	0.3688	0.0567	0.4713	0.2238	0.3159	0.4666			
IBEX	0.0727	0.0024	-0.0605	0.9881	0.1558	1.3298	1.1022	0.9881	57.7845
p-val	0.0047	0.2396	0.0004	0.0000	0.0000	0.0000			
λ	0.5135	0.1643	0.4112	0.1913	0.3767	0.3921			
SMI	0.0461	-0.0031	-0.0726	0.9862	0.1453	1.4407	0.8939	0.9862	50.0225
p-val	0.0103	0.1266	0.0000	0.0000	0.0000	0.0000			
λ	0.3795	0.0554	0.2723	0.1695	0.2520	0.4583			
DAX	0.0919	0.0041	-0.0573	0.9871	0.1213	1.2412	1.1690	0.9871	53.5189
p-val	0.0015	0.1076	0.0007	0.0000	0.0000	0.0000			
λ	0.6011	0.1706	0.4521	0.1759	0.3661	0.6497			
CAC	0.0214	0.0030	-0.0588	0.9903	0.1169	1.3236	1.1650	0.9903	71.3910
p-val	0.1813	0.1491	0.0001	0.0000	0.0000	0.0000			
λ	0.2033	0.0958	0.3285	0.1329	0.3053	0.4313			

CUADRO 4.14: Resultados del modelo de Nelson [1991] para índices de mercado, porcentaje de NA=30%

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
IBM	0.0267	0.0359	0.0874	0.9038	0.3543	0.1138	1.2593	1.2870	0.9720	24.4293
p-val	0.1203	0.0020	0.0000	0.0000	0.0405	0.2904	0.0000			
λ	0.1119	0.1672	0.1846	0.1722	0.6402	0.9018	0.1663			
APPLE	0.1930	0.0641	0.0700	0.9179	0.3571	0.0654	1.2865	2.2863	0.9719	24.3247
p-val	0.0000	0.0120	0.0000	0.0000	0.0401	0.3867	0.0000			
λ	0.1201	0.3271	0.3230	0.4937	0.6542	0.9182	0.0823			
GE	0.0050	0.0093	0.0643	0.9464	0.4012	-0.0080	1.3255	1.6706	0.9944	124.0891
p-val	0.4072	0.0177	0.0000	0.0000	0.0016	0.4704	0.0000			
λ	0.1629	0.1531	0.2199	0.2245	0.2043	0.2428	0.1593			
TEF	0.0038	0.0307	0.1001	0.8946	0.1225	0.3598	1.3183	1.4393	0.9786	32.0996
p-val	0.4017	0.0001	0.0000	0.0000	0.3096	0.1041	0.0000			
λ	0.1233	0.1369	0.4264	0.5769	0.8515	0.9577	0.4115			
SAN	0.0106	0.0190	0.0806	0.9372	0.7485	-0.1197	1.2579	1.9782	0.9904	71.4889
p-val	0.3174	0.0008	0.0000	0.0000	0.0000	0.1928	0.0000			
λ	0.1344	0.1648	0.2798	0.3202	0.3575	0.7207	0.5431			
NOVARTIS	0.0390	0.0190	0.0598	0.9512	0.6454	-0.3584	1.3075	1.2343	0.9846	44.6856
p-val	0.0361	0.0124	0.0000	0.0000	0.0101	0.1323	0.0000			
λ	0.1072	0.0824	0.1269	0.2739	0.7236	0.9587	0.2065			
BMW	0.0356	0.0168	0.0695	0.9373	0.1774	0.1675	1.3781	1.9428	0.9913	79.0820
p-val	0.1244	0.0089	0.0000	0.0000	0.2330	0.3238	0.0000			
λ	0.1150	0.2352	0.1986	0.2371	0.8398	0.9583	0.1624			
EAD	0.0414	0.0298	0.0653	0.9353	0.4186	0.0784	1.3859	2.3047	0.9870	53.1658
p-val	0.1465	0.0026	0.0000	0.0000	0.0059	0.2815	0.0000			
λ	0.1229	0.1048	0.1910	0.2653	0.3765	0.7640	0.2377			

CUADRO 4.15: Resultados del modelo de Hentschel [1995] para firmas cotizadas, porcentaje de NA=10%

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
IBM	0.0283	0.0310	0.0799	0.9154	0.2815	0.0477	1.2689	1.2680	0.9755	27.9796
p-val	0.1174	0.0025	0.0000	0.0000	0.0713	0.3972	0.0000			
λ	0.1859	0.1742	0.1959	0.2151	0.6335	0.8826	0.2890			
APPLE	0.1894	0.0409	0.0573	0.9428	0.4490	-0.1590	1.3024	2.2980	0.9822	38.5465
p-val	0.0000	0.0185	0.0000	0.0000	0.0062	0.1824	0.0000			
λ	0.0680	0.2143	0.2867	0.3714	0.5730	0.8116	0.4243			
GE	0.0020	0.0078	0.0590	0.9467	0.2979	0.1896	1.3350	1.6464	0.9952	145.1064
p-val	0.4635	0.0114	0.0000	0.0000	0.0480	0.1708	0.0000			
λ	0.1796	0.0751	0.2717	0.2923	0.5708	0.8953	0.4635			
TEF	0.0136	0.0353	0.0971	0.8952	0.1963	0.2421	1.3003	1.4184	0.9750	27.4331
p-val	0.2738	0.0003	0.0000	0.0000	0.0893	0.0344	0.0000			
λ	0.1568	0.2124	0.2454	0.3251	0.5449	0.7781	0.3895			
SAN	0.0173	0.0199	0.0892	0.9357	0.7598	-0.2175	1.3005	1.9036	0.9895	65.9465
p-val	0.2535	0.0009	0.0000	0.0000	0.0000	0.1077	0.0000			
λ	0.1798	0.1884	0.1908	0.4363	0.4946	0.8787	0.2711			
NOVARTIS	0.0229	0.0138	0.0511	0.9581	0.5844	-0.2740	1.3002	1.1909	0.9884	59.3714
p-val	0.1616	0.0214	0.0001	0.0000	0.0382	0.2580	0.0000			
λ	0.2120	0.1669	0.1227	0.3874	0.8077	0.9682	0.2447			
BMW	0.0311	0.0148	0.0628	0.9447	0.2705	0.0218	1.3904	1.8867	0.9921	87.7305
p-val	0.1616	0.0117	0.0000	0.0000	0.0641	0.4526	0.0000			
λ	0.1397	0.2320	0.1846	0.2133	0.6396	0.8866	0.2042			
EAD	0.0533	0.0305	0.0553	0.9468	0.5367	-0.0580	1.3449	2.2519	0.9864	50.7524
p-val	0.0990	0.0039	0.0000	0.0000	0.0039	0.3986	0.0000			
λ	0.1704	0.2160	0.2026	0.3012	0.5071	0.8979	0.3806			

CUADRO 4.16: Resultados del modelo de Hentschel [1995] para firmas cotizadas, porcentaje de NA=15%

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
IBM	0.0209	0.0228	0.0658	0.9335	0.2098	0.0090	1.2094	1.2497	0.9818	37.6519
p-val	0.2251	0.0191	0.0002	0.0000	0.1484	0.4775	0.0000			
λ	0.2566	0.2593	0.3497	0.3609	0.6377	0.8713	0.4614			
APPLE	0.1308	0.0373	0.0552	0.9459	0.4387	-0.1334	1.2803	2.4347	0.9845	44.3555
p-val	0.0050	0.0472	0.0001	0.0000	0.0116	0.2764	0.0000			
λ	0.3914	0.3735	0.3367	0.4999	0.6078	0.9330	0.2825			
GE	0.0112	0.0067	0.0539	0.9571	0.3299	-0.0660	1.2957	1.5887	0.9958	164.4605
p-val	0.3310	0.0360	0.0000	0.0000	0.0801	0.4033	0.0000			
λ	0.2431	0.1477	0.2597	0.3216	0.7018	0.9454	0.4818			
TEF	0.0076	0.0282	0.0854	0.9213	0.3993	-0.1255	1.2934	1.3677	0.9794	33.2212
p-val	0.3568	0.0038	0.0000	0.0000	0.0891	0.3661	0.0000			
λ	0.3387	0.2429	0.1470	0.4759	0.8511	0.9745	0.3941			
SAN	0.0105	0.0136	0.0793	0.9365	0.4465	-0.0284	1.2638	1.9189	0.9929	96.8636
p-val	0.3219	0.0053	0.0000	0.0000	0.0482	0.4672	0.0000			
λ	0.2797	0.2449	0.3300	0.5943	0.8593	0.9776	0.5441			
NOVARTIS	0.0302	0.0131	0.0558	0.9617	0.6388	-0.5728	1.2700	1.2263	0.9893	64.3533
p-val	0.1035	0.0094	0.0024	0.0000	0.0031	0.0631	0.0000			
λ	0.2872	0.2023	0.7567	0.2739	0.7487	0.9467	0.3863			
BMW	0.0688	0.0122	0.0606	0.9481	0.1497	0.0742	1.4092	1.8733	0.9934	105.1128
p-val	0.0306	0.0117	0.0000	0.0000	0.2767	0.4115	0.0000			
λ	0.2888	0.1630	0.3645	0.2861	0.8489	0.9608	0.1626			
EAD	0.0481	0.0245	0.0613	0.9415	0.0767	0.2181	1.3525	2.3757	0.9896	66.5214
p-val	0.1431	0.0104	0.0000	0.0000	0.3662	0.2062	0.0000			
λ	0.3023	0.2029	0.2266	0.2912	0.6218	0.8577	0.5178			

CUADRO 4.17: Resultados del modelo de Hentschel [1995] para firmas cotizadas, porcentaje de NA=30 %

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0604	0.0193	0.0769	0.9438	0.7653	-0.3577	1.4545	1.1278	0.9830	40.4249
p-val	0.0044	0.0019	0.0000	0.0000	0.0024	0.2025	0.0000			
λ	0.1905	0.3937	0.4715	0.8138	0.8362	0.9841	0.3458			
SP500	0.0550	0.0125	0.0730	0.9549	0.9184	-0.4573	1.2782	0.8654	0.9856	47.8998
p-val	0.0008	0.0004	0.0000	0.0000	0.0000	0.1049	0.0000			
λ	0.1909	0.2174	0.5390	0.6329	0.4374	0.9532	0.3669			
EUROSTOXX	0.0185	0.0181	0.0658	0.9192	0.8433	0.2737	1.4891	1.3633	0.9867	51.7952
p-val	0.1874	0.0003	0.0000	0.0000	0.0008	0.1114	0.0000			
λ	0.0736	0.3576	0.3146	0.6870	0.5328	0.8761	0.3457			
IBEX	0.0537	0.0145	0.0634	0.9373	0.7893	0.0556	1.3886	1.1695	0.9877	55.7979
p-val	0.0101	0.0002	0.0000	0.0000	0.0011	0.4074	0.0000			
λ	0.2668	0.1949	0.2201	0.6544	0.6294	0.9025	0.2651			
SMI	0.0288	0.0175	0.0658	0.9306	0.8981	0.0277	1.4778	0.9781	0.9821	38.3579
p-val	0.0523	0.0000	0.0000	0.0000	0.0001	0.4559	0.0000			
λ	0.1734	0.1201	0.2037	0.5691	0.5207	0.7791	0.5395			
DAX	0.0461	0.0192	0.0692	0.9161	0.5112	0.3719	1.3611	1.2669	0.9848	45.3801
p-val	0.0211	0.0002	0.0000	0.0000	0.0374	0.0886	0.0000			
λ	0.2620	0.2973	0.3555	0.7341	0.7331	0.9600	0.2820			
CAC	0.0157	0.0175	0.0645	0.9319	0.8957	0.0866	1.5195	1.3027	0.9865	51.1308
p-val	0.2274	0.0002	0.0000	0.0000	0.0000	0.2928	0.0000			
λ	0.0955	0.2448	0.2123	0.5281	0.4155	0.8571	0.2820			

CUADRO 4.18: Resultados del modelo de Hentschel [1995] para índices de mercado, porcentaje de NA=10%

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0758	0.0209	0.0790	0.9263	0.4096	0.2168	1.3645	1.1941	0.9825	39.1693
p-val	0.0004	0.0004	0.0129	0.0000	0.2292	0.3675	0.0000			
λ	0.1324	0.2794	0.9017	0.8905	0.9527	0.9864	0.4246			
SP500	0.0530	0.0144	0.0739	0.9388	0.8055	-0.1641	1.2793	0.9213	0.9844	44.0402
p-val	0.0020	0.0003	0.0000	0.0000	0.0006	0.2780	0.0000			
λ	0.3038	0.1500	0.2504	0.5877	0.5893	0.9492	0.3304			
EUROSTOXX	0.0133	0.0188	0.0714	0.9243	0.7431	0.1208	1.4407	1.2656	0.9852	46.4567
p-val	0.2741	0.0002	0.0000	0.0000	0.0007	0.3010	0.0000			
λ	0.1913	0.2428	0.2101	0.6284	0.5634	0.8905	0.2754			
IBEX	0.0325	0.0159	0.0640	0.9388	0.8253	-0.0103	1.4160	1.1718	0.9864	50.5831
p-val	0.0693	0.0003	0.0000	0.0000	0.0002	0.4804	0.0000			
λ	0.2093	0.2670	0.3084	0.6038	0.5566	0.8840	0.5011			
SMI	0.0254	0.0179	0.0726	0.9215	0.6867	0.1022	1.5079	0.9796	0.9817	37.5531
p-val	0.0828	0.0000	0.0000	0.0000	0.0035	0.3232	0.0000			
λ	0.1977	0.1992	0.3616	0.6121	0.6932	0.9254	0.3671			
DAX	0.0618	0.0178	0.0659	0.9265	0.6124	0.2119	1.3416	1.2191	0.9854	46.9720
p-val	0.0051	0.0002	0.0000	0.0000	0.0090	0.1740	0.0000			
λ	0.2987	0.2063	0.3408	0.6139	0.6395	0.9170	0.2922			
CAC	0.0191	0.0177	0.0696	0.9255	0.6895	0.1776	1.5161	1.2855	0.9862	49.9849
p-val	0.2012	0.0004	0.0000	0.0000	0.0266	0.3027	0.0000			
λ	0.2365	0.3347	0.5685	0.8163	0.8299	0.9629	0.3168			

CUADRO 4.19: Resultados del modelo de Hentschel [1995] para índices de mercado, porcentaje de NA=15%

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0495	0.0157	0.0620	0.9406	0.3523	0.0272	1.3503	1.1871	0.9868	52.2278
p-val	0.0393	0.0039	0.0000	0.0000	0.1032	0.4604	0.0000			
λ	0.4944	0.2206	0.2817	0.4646	0.7583	0.9177	0.5371			
SP500	0.0539	0.0096	0.0758	0.9564	0.7944	-0.5806	1.2234	0.8226	0.9883	59.1155
p-val	0.0035	0.0052	0.0000	0.0000	0.0013	0.0410	0.0000			
λ	0.3377	0.3001	0.3752	0.6229	0.7570	0.9417	0.3504			
EUROSTOXX	0.0011	0.0199	0.0803	0.9163	0.4538	0.1934	1.3857	1.3339	0.9851	46.1411
p-val	0.4812	0.0026	0.0000	0.0000	0.0172	0.1701	0.0000			
λ	0.2661	0.3147	0.4037	0.5490	0.6249	0.8622	0.3526			
IBEX	0.0573	0.0122	0.0647	0.9404	0.5502	0.0815	1.3456	1.2180	0.9901	69.5707
p-val	0.0577	0.0041	0.0001	0.0000	0.0561	0.4009	0.0000			
λ	0.6940	0.4108	0.6054	0.7426	0.8451	0.9598	0.4309			
SMI	0.0350	0.0148	0.0785	0.9234	0.3282	0.1014	1.3855	0.9761	0.9848	45.1201
p-val	0.0435	0.0026	0.0000	0.0000	0.0851	0.3596	0.0000			
λ	0.3791	0.2920	0.3036	0.5224	0.7632	0.9133	0.5410			
DAX	0.0782	0.0204	0.0847	0.9149	0.3069	0.2207	1.4353	1.2250	0.9832	41.0066
p-val	0.0464	0.0013	0.0000	0.0000	0.1918	0.3180	0.0000			
λ	0.8034	0.4217	0.5290	0.7641	0.9113	0.9866	0.8430			
CAC	0.0306	0.0115	0.0654	0.9441	0.5500	-0.1097	1.4044	1.1520	0.9900	69.2617
p-val	0.0998	0.0029	0.0000	0.0000	0.0039	0.1589	0.0000			
λ	0.2534	0.1553	0.3705	0.3378	0.5909	0.6784	0.5408			

CUADRO 4.20: Resultados del modelo de Hentschel [1995] para índices de mercado, porcentaje de NA=30%

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
IBM	0.0365	0.0164	0.0644	0.8064	0.9907	0.0347	1.2665	1.1518	0.9907	73.8723
p-val	0.0509	0.0084	0.0017	0.0000	0.0000	0.0005	0.0000			
λ	0.0864	0.0309	0.1368	0.0712	0.0215	0.0314	0.2843			
APPLE	0.1700	0.0234	0.0609	0.6537	0.9958	0.0221	1.2295	1.5371	0.9958	164.9253
p-val	0.0000	0.0506	0.0125	0.0061	0.0000	0.0002	0.0000			
λ	0.0946	0.0460	0.1630	0.2715	0.0463	0.0828	0.2536			
GE	0.0194	0.0078	0.0502	0.7272	0.9979	0.0398	1.2921	1.3867	0.9979	325.2568
p-val	0.2294	0.0161	0.0227	0.0007	0.0000	0.0000	0.0000			
λ	0.1647	0.0312	0.2186	0.2130	0.0265	0.0773	0.2632			
TEF	0.0344	0.0082	0.0741	0.8850	0.9965	0.0211	1.3606	1.2367	0.9965	196.2832
p-val	0.0961	0.1338	0.0000	0.0000	0.0000	0.1099	0.0000			
λ	0.2296	0.0451	0.0582	0.1838	0.0517	0.0446	0.2525			
SAN	0.0509	0.0078	0.0643	0.8677	0.9988	0.0390	1.2278	1.6001	0.9988	578.8353
p-val	0.0716	0.0608	0.0006	0.0000	0.0000	0.0175	0.0000			
λ	0.2340	0.0688	0.1327	0.2327	0.0218	0.1554	0.3412			
NOVARTIS	0.0537	0.0135	0.0778	0.6062	0.9912	0.0289	1.3179	1.1135	0.9912	78.3648
p-val	0.0078	0.0114	0.0031	0.0001	0.0000	0.0001	0.0000			
λ	0.1294	0.0544	0.1430	0.1646	0.0527	0.0697	0.2240			
BMW	0.0559	0.0098	0.0372	0.9107	0.9982	0.0312	1.3623	1.5334	0.9982	377.7087
p-val	0.0386	0.0570	0.0160	0.0000	0.0000	0.0044	0.0000			
λ	0.1583	0.0445	0.1178	0.1905	0.0590	0.1174	0.2659			
EAD	0.0975	0.0370	0.0599	0.8151	0.9946	0.0364	1.3208	1.6206	0.9946	127.9024
p-val	0.0069	0.0789	0.0075	0.0000	0.0000	0.0337	0.0000			
λ	0.1250	0.1024	0.1001	0.3543	0.0876	0.1992	0.2683			

CUADRO 4.21: Resultados del modelo de Lee% Engle [1999] para firmas cotizadas, porcentaje de NA=10%

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
IBM	0.0601	0.0166	0.0463	0.8097	0.9904	0.0360	1.2755	1.1453	0.9904	71.5361
p-val	0.0031	0.0058	0.0172	0.0000	0.0000	0.0001	0.0000			
λ	0.1343	0.0610	0.1687	0.1094	0.0552	0.0863	0.2475			
APPLE	0.1821	0.0321	0.0607	0.6074	0.9943	0.0251	1.2392	1.5418	0.9943	121.5686
p-val	0.0000	0.0351	0.0139	0.0006	0.0000	0.0002	0.0000			
λ	0.2536	0.1108	0.2826	0.1745	0.0913	0.1581	0.2804			
GE	0.0175	0.0098	0.0561	0.7475	0.9972	0.0427	1.3044	1.3735	0.9972	248.4497
p-val	0.2651	0.0213	0.0151	0.0000	0.0000	0.0000	0.0000			
λ	0.3351	0.1534	0.2180	0.2267	0.0861	0.1308	0.4262			
TEF	0.0262	0.0100	0.0680	0.8944	0.9956	0.0185	1.2938	1.2292	0.9956	156.1411
p-val	0.1712	0.1724	0.0007	0.0000	0.0000	0.1853	0.0000			
λ	0.2947	0.0778	0.1090	0.1574	0.0777	0.0823	0.4318			
SAN	0.0236	0.0083	0.0709	0.8795	0.9986	0.0346	1.2407	1.5451	0.9986	480.3094
p-val	0.2078	0.0886	0.0003	0.0000	0.0000	0.0548	0.0000			
λ	0.1712	0.1027	0.1044	0.2196	0.0254	0.1541	0.1616			
NOVARTIS	0.0464	0.0122	0.0616	0.4990	0.9922	0.0281	1.2970	1.1198	0.9922	88.8374
p-val	0.0255	0.0129	0.0270	0.0092	0.0000	0.0000	0.0000			
λ	0.2101	0.0626	0.3431	0.0601	0.0560	0.0683	0.3401			
BMW	0.0545	0.0120	0.0196	0.5085	0.9985	0.0453	1.3688	1.7039	0.9985	473.7217
p-val	0.0417	0.0530	0.2165	0.2744	0.0000	0.0007	0.0000			
λ	0.1089	0.2002	0.2366	0.2349	0.1031	0.4654	0.1829			
EAD	0.0583	0.0251	0.0748	0.7062	0.9961	0.0354	1.4537	1.5984	0.9961	175.6537
p-val	0.0758	0.0744	0.0081	0.0002	0.0000	0.0027	0.0000			
λ	0.1438	0.1672	0.3827	0.4668	0.1474	0.2781	0.3975			

CUADRO 4.22: Resultados del modelo de Lee% Engle [1999] para firmas cotizadas, porcentaje de NA=15%

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
IBM	0.0396	0.0129	0.0486	0.8385	0.9925	0.0230	1.2315	1.1427	0.9925	91.5419
p-val	0.0800	0.0309	0.0129	0.0000	0.0000	0.0056	0.0000			
λ	0.3570	0.0821	0.3157	0.1759	0.0717	0.1117	0.3773			
APPLE	0.2033	0.0378	0.0356	0.8192	0.9928	0.0225	1.2329	1.5134	0.9928	95.4587
p-val	0.0000	0.0529	0.0330	0.0002	0.0000	0.0122	0.0000			
λ	0.3861	0.1351	0.2629	0.3242	0.1429	0.2266	0.3344			
GE	0.0180	0.0090	0.0353	0.4829	0.9969	0.0385	1.2645	1.3049	0.9969	221.2148
p-val	0.2557	0.0186	0.1331	0.1346	0.0000	0.0000	0.0000			
λ	0.2994	0.1203	0.3219	0.1859	0.0740	0.2321	0.3554			
TEF	0.0193	0.0222	0.0362	0.8118	0.9897	0.0349	1.2938	1.2109	0.9897	67.2347
p-val	0.2353	0.2474	0.2206	0.0284	0.0000	0.2291	0.0000			
λ	0.3741	0.1318	0.3067	0.4008	0.1159	0.3328	0.5908			
SAN	0.0262	0.0065	0.0506	0.9084	0.9988	0.0317	1.2146	1.5301	0.9988	584.3337
p-val	0.2150	0.0929	0.0020	0.0000	0.0000	0.0386	0.0000			
λ	0.3641	0.1321	0.1503	0.2861	0.0290	0.2361	0.5151			
NOVARTIS	0.0390	0.0179	0.0346	0.6575	0.9878	0.0353	1.4106	1.1014	0.9878	56.3515
p-val	0.0639	0.0076	0.2347	0.0506	0.0000	0.1788	0.0000			
λ	0.3538	0.1067	0.2320	0.3089	0.1260	0.0248	0.2874			
BMW	0.0543	0.0085	0.0154	0.3968	0.9989	0.0353	1.3321	1.7066	0.9989	635.9849
p-val	0.0613	0.0575	0.2624	0.4362	0.0000	0.0000	0.0000			
λ	0.2510	0.1646	0.3911	0.0242	0.1465	0.2559	0.4484			
EAD	0.0629	0.0426	0.0390	0.5450	0.9932	0.0452	1.3281	1.6087	0.9932	102.2625
p-val	0.0762	0.0387	0.1146	0.1641	0.0000	0.0016	0.0000			
λ	0.3070	0.3482	0.3943	0.3420	0.2960	0.4024	0.6575			

CUADRO 4.23: Resultados del modelo de Lee% Engle [1999] para firmas cotizadas, porcentaje de NA=30%

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0769	0.0206	0.0052	0.9705	0.9879	0.0623	1.4217	1.1422	0.9879	56.9318
p-val	0.0002	0.0009	0.4849	0.0000	0.0000	0.3256	0.0000			
λ	0.0727	0.0376	0.0044	0.0456	0.0291	0.0046	0.1760			
SP500	0.0697	0.0096	0.0110	0.7787	0.9930	0.0614	1.3089	1.0876	0.9930	98.5399
p-val	0.0000	0.0167	0.4676	0.1073	0.0000	0.3249	0.0000			
λ	0.0783	0.3729	0.0247	0.3599	0.2218	0.0236	0.2259			
EUROSTOXX	0.0555	0.0055	0.0559	0.9150	0.9974	0.0296	1.3917	1.2084	0.9974	267.6843
p-val	0.0141	0.0653	0.0012	0.0000	0.0000	0.0558	0.0000			
λ	0.3290	0.0208	0.0405	0.0749	0.0146	0.0423	0.2962			
IBEX	0.0708	0.0035	0.0621	0.8741	0.9984	0.0250	1.3733	1.2053	0.9984	434.0467
p-val	0.0020	0.1134	0.0014	0.0002	0.0000	0.0643	0.0000			
λ	0.4125	0.3487	0.4927	0.4882	0.1238	0.3747	0.1748			
SMI	0.0607	0.0044	0.0672	0.8945	0.9960	0.0249	1.4161	1.0285	0.9960	174.8501
p-val	0.0003	0.0589	0.0001	0.0000	0.0000	0.0432	0.0000			
λ	0.1679	0.0917	0.0503	0.0588	0.0939	0.0340	0.3341			
DAX	0.0889	0.0071	0.0311	0.8506	0.9978	0.0449	1.3147	1.3865	0.9978	313.3183
p-val	0.0001	0.0785	0.1772	0.0159	0.0000	0.0815	0.0000			
λ	0.2238	0.3184	0.1743	0.2979	0.2413	0.1469	0.4698			
CAC	0.0600	0.0052	0.0532	0.8863	0.9973	0.0249	1.3977	1.1766	0.9973	256.5199
p-val	0.0022	0.0939	0.0045	0.1250	0.0000	0.0937	0.0000			
λ	0.0805	0.1906	0.3828	0.0480	0.0626	0.2943	0.3486			

CUADRO 4.24: Resultados del modelo de Lee% Engle [1999] índices de mercado, porcentaje de NA=10%

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0656	0.0141	0.0121	0.9290	0.9921	0.0437	1.3326	1.1590	0.9921	87.3432
p-val	0.0032	0.1602	0.4353	0.0000	0.0000	0.2777	0.0000			
λ	0.2560	0.0204	0.0428	0.3303	0.0305	0.0368	0.4007			
SP500	0.0677	0.0069	0.0265	0.6759	0.9953	0.0358	1.1888	1.0958	0.9953	147.3613
p-val	0.0005	0.0925	0.3061	0.1822	0.0000	0.2490	0.0000			
λ	0.4103	0.3116	0.2338	0.2778	0.1428	0.2148	0.2773			
EUROSTOXX	0.0500	0.0063	0.0449	0.8593	0.9975	0.0372	1.3812	1.2550	0.9975	275.3650
p-val	0.0149	0.1042	0.0989	0.0200	0.0000	0.1434	0.0000			
λ	0.2431	0.3806	0.2531	0.3433	0.1801	0.2215	0.3408			
IBEX	0.0804	0.0033	0.0555	0.8817	0.9985	0.0252	1.3201	1.2007	0.9985	454.8139
p-val	0.0002	0.1162	0.0022	0.0007	0.0000	0.0639	0.0000			
λ	0.3643	0.3911	0.4892	0.4068	0.0876	0.4395	0.3274			
SMI	0.0550	0.0048	0.0471	0.9192	0.9958	0.0289	1.3921	1.0411	0.9958	164.0582
p-val	0.0007	0.0506	0.0063	0.0000	0.0000	0.0527	0.0000			
λ	0.1416	0.0290	0.0847	0.1292	0.0590	0.0760	0.1737			
DAX	0.0874	0.0078	0.0356	0.8603	0.9972	0.0411	1.3096	1.2866	0.9972	244.6241
p-val	0.0001	0.0970	0.2031	0.0131	0.0000	0.1694	0.0000			
λ	0.1721	0.3469	0.2071	0.1898	0.1109	0.1960	0.3088			
CAC	0.0524	0.0056	0.0445	0.8597	0.9971	0.0259	1.3946	1.1654	0.9971	237.2303
p-val	0.0161	0.1406	0.0340	0.0014	0.0000	0.1409	0.0000			
λ	0.3311	0.3419	0.5378	0.4525	0.1448	0.4364	0.2982			

CUADRO 4.25: Resultados del modelo de Lee% Engle [1999] índices de mercado, porcentaje de NA=15%

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
NASDAQ	0.0717	0.0149	0.0030	0.7138	0.9920	0.0517	1.3272	1.1711	0.9920	86.4114
p-val	0.0119	0.0061	0.4616	0.1089	0.0000	0.0490	0.0000			
λ	0.5860	0.0848	0.0406	0.2045	0.0766	0.0246	0.3451			
SP500	0.0758	0.0124	0.0072	0.6282	0.9916	0.0622	1.2148	1.1016	0.9916	82.2020
p-val	0.0000	0.0098	0.4752	0.2295	0.0000	0.2960	0.0000			
λ	0.1401	0.2723	0.0243	0.2193	0.1248	0.0252	0.4141			
EUROSTOXX	0.0367	0.0051	0.0451	0.9253	0.9979	0.0326	1.3596	1.2657	0.9979	336.2345
p-val	0.0822	0.0813	0.0140	0.0000	0.0000	0.0647	0.0000			
λ	0.4109	0.0248	0.0846	0.2239	0.0376	0.0502	0.3074			
IBEX	0.0924	0.0038	0.0480	0.8768	0.9986	0.0266	1.1947	1.2661	0.9986	497.0037
p-val	0.0001	0.1141	0.0067	0.0000	0.0000	0.0511	0.0000			
λ	0.3349	0.3817	0.2973	0.3812	0.0577	0.5233	0.6616			
SMI	0.0425	0.0061	0.0539	0.8742	0.9950	0.0361	1.3660	1.0780	0.9950	138.6271
p-val	0.0174	0.0316	0.0041	0.0000	0.0000	0.0071	0.0000			
λ	0.3638	0.1180	0.1819	0.2336	0.1833	0.1372	0.4740			
DAX	0.0903	0.0110	0.0141	0.5396	0.9967	0.0570	1.3369	1.3730	0.9967	211.4704
p-val	0.0021	0.0439	0.2866	0.2636	0.0000	0.0067	0.0000			
λ	0.5377	0.1770	0.5058	0.1713	0.0936	0.5222	0.6171			
CAC	0.0648	0.0084	0.0289	0.6301	0.9964	0.0416	1.4008	1.2127	0.9964	190.5553
p-val	0.0065	0.1098	0.1743	0.1737	0.0000	0.0860	0.0000			
λ	0.3780	0.5665	0.8027	0.3193	0.2636	0.7329	0.4925			

CUADRO 4.26: Resultados del modelo de Lee & Engle [1999] índices de mercado, porcentaje de NA=30%

Anexo II: Cuadros comparativos

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
IBM	1.1870	1.7687	1.3985	0.9894	1.4338	1.0207	0.9992	0.9894	0.5609
APPLE	1.0148	1.4616	1.4641	0.9903	1.2664	1.0246	0.9831	0.9903	0.6681
GE	0.8036	1.2549	1.2187	0.9985	1.1504	1.0515	0.9947	0.9985	0.7845
TEF	0.9526	0.9813	1.2235	0.9989	1.0450	1.0273	0.9861	0.9989	0.9605
SAN	0.5409	1.0853	1.2068	0.9977	1.0856	1.0384	0.9684	0.9977	0.8699
NOVARTIS	**1.0472**	0.9678	1.4013	0.9994	0.9665	1.0194	0.9906	0.9994	0.9553
BMW	1.1647	1.0978	1.0660	0.9988	1.0721	0.9876	0.9888	0.9988	0.8929
EAD	**0.9412**	0.9667	1.2532	0.9999	1.0780	1.0927	0.9704	0.9999	0.9941
NASDAQ	0.8346	1.3078	1.3515	0.9956	1.0278	1.0485	0.9970	0.9956	0.7247
SP500	1.0019	1.4337	1.3375	0.9962	1.0919	1.0895	0.9997	0.9962	0.7125
EUROSTOXX	0.6598	1.1340	1.5194	0.9949	0.8368	1.1243	0.9811	0.9949	0.7085
IBEX	1.2322	2.6067	1.1845	0.9958	1.1501	1.0672	1.0238	0.9958	0.7369
SMI	1.0915	1.0278	1.1675	1.0006	1.0686	1.0690	0.9917	1.0006	1.0307
DAX	0.8903	0.7055	1.3371	0.9972	0.9646	1.0745	0.9687	0.9972	0.8308
CAC	0.7887	1.0379	1.3471	0.9968	0.8707	1.1141	0.9861	0.9968	0.8306

CUADRO 4.27: Comparación de los coeficientes estimados para el modelo de Nelson [1991], porcentaje de NA=10 %

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
IBM	**0.9603**	1.5358	1.4967	0.9909	1.2811	1.0281	0.9934	0.9909	0.6245
APPLE	1.2842	0.9626	1.2261	1.0004	1.0435	1.0475	0.9811	1.0004	1.0146
GE	0.2067	1.5182	1.0512	0.9981	1.2162	1.0108	1.0323	0.9981	0.7252
TEF	0.5706	0.9570	1.2342	0.9963	1.1187	1.0483	0.9510	0.9963	0.8631
SAN	1.6362	1.3111	1.3098	0.9946	1.1909	1.1237	0.9448	0.9946	0.6884
NOVARTIS	**0.7504**	**0.9353**	1.0433	1.0001	1.0466	1.0465	0.9930	1.0001	1.0096
BMW	**0.7245**	1.2118	1.1347	0.9981	1.0971	0.9908	0.9991	0.9981	0.8231
EAD	**0.7184**	1.0962	1.5065	0.9984	1.1654	1.0673	0.9827	0.9984	0.8903
NASDAQ	0.9425	1.6600	1.7090	0.9938	1.1963	1.0942	1.0034	0.9938	0.6074
SP500	0.9660	1.6423	1.5671	0.9946	0.9876	1.1095	1.0087	0.9946	0.5960
EUROSTOXX	0.4340	1.5756	1.6843	0.9935	0.9908	1.1387	0.9996	0.9935	0.6274
IBEX	1.2896	4.0813	1.3634	0.9951	1.2044	1.0588	1.0337	0.9951	0.6939
SMI	**1.8824**	1.3335	1.4233	0.9939	1.1986	1.1132	1.0117	0.9939	0.6955
DAX	0.7987	0.6527	1.5203	0.9968	1.0608	1.1089	0.9607	0.9968	0.8065
CAC	**0.6007**	1.1129	1.4478	0.9944	0.8510	1.0736	0.9759	0.9944	0.7061

CUADRO 4.28: Comparación de los coeficientes estimados para el modelo de Nelson [1991], porcentaje de NA=15 %

	μ	ω	α	β	τ	ψ	σ_y	ϕ	$h2l$
IBM	**0.6761**	2.5654	3.3210	0.9850	1.5259	1.0471	0.9979	0.9850	0.3802
APPLE	1.0756	1.4788	1.6340	0.9897	1.3299	1.0122	0.9739	0.9897	0.6508
GE	-0.6745	1.7469	1.5006	0.9965	1.2453	1.1348	0.9195	0.9965	0.4790
TEF	0.9270	1.5605	1.4692	0.9896	1.3910	1.0605	0.9897	0.9896	0.6108
SAN	3.5058	1.8491	3.0721	0.9907	1.1952	1.2106	0.9241	0.9907	0.4665
NOVARTIS	**0.5879**	0.9672	1.9723	0.9986	1.0640	1.0640	0.9823	0.9986	0.8908
BMW	**0.9061**	1.5545	2.8463	0.9959	1.2258	1.0637	0.9881	0.9959	0.6286
EAD	**0.6282**	1.4787	2.0628	0.9945	1.3634	1.0626	0.9444	0.9945	0.6270
NASDAQ	0.8925	1.0810	2.5253	0.9931	1.0174	1.1310	0.9510	0.9931	0.5622
SP500	0.7832	**3.5958**	2.4991	0.9958	0.8179	1.1617	0.8793	0.9958	0.6823
EUROSTOXX	0.5230	1.3618	2.4247	0.9911	0.9705	1.2156	0.9603	0.9911	0.4926
IBEX	0.7214	0.6476	1.9034	0.9961	0.7890	1.1312	0.9525	0.9961	0.7529
SMI	0.6204	**1.8714**	1.8737	0.9939	0.7959	1.1275	0.9670	0.9939	0.6932
DAX	0.7255	0.3651	2.1937	0.9965	0.9919	1.1866	0.8954	0.9965	0.7887
CAC	0.8233	1.1383	2.6463	0.9906	0.8937	1.2362	0.9387	0.9906	0.5068

CUADRO 4.29: Comparación de los coeficientes estimados para el modelo de Nelson [1991], porcentaje de NA=30 %

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
IBM	1.0990	1.1030	1.1164	0.9879	1.0360	0.5043	1.0482	0.9720	0.9962	0.8821
APPLE	0.8877	1.2521	1.2853	0.9608	**0.1515**	**8.3658**	1.0127	0.9993	0.9927	0.7966
GE	0.0235	1.3764	1.1213	0.9896	1.0494	**-4.2990**	1.0282	0.9608	0.9976	0.7003
TEF	0.8935	1.1690	1.2554	0.9628	-0.2718	1.7326	1.0510	1.0424	0.9974	0.8922
SAN	1.6177	1.4124	1.2391	0.9909	1.1575	**2.2041**	1.1126	0.9044	0.9946	0.6412
NOVARTIS	0.9072	0.9446	0.9837	1.0069	1.2878	**1.3476**	1.0344	0.9666	1.0004	1.0234
BMW	1.1754	1.3303	1.1287	0.9916	**1.7360**	-0.2856	1.0354	0.9628	0.9967	0.7280
EAD	1.2549	1.1497	1.1023	0.9897	0.8467	**1.8548**	1.0054	0.9765	0.9977	0.8487
NASDAQ	0.9241	1.2086	0.8866	0.9965	1.3067	**0.6150**	1.0433	1.0043	0.9963	0.8235
SP500	0.8319	1.3943	0.9613	0.9728	1.0888	**0.0124**	1.0828	1.0976	0.9960	0.7844
EUROSTOXX	0.0907	1.3223	1.0195	0.9821	1.1858	**1.2899**	1.0834	0.9896	0.9955	0.7484
IBEX	0.8766	1.3388	1.0952	0.9989	1.2670	-1.7975	1.0913	0.9754	0.9953	0.7262
SMI	0.7767	1.2315	1.0622	0.9689	0.9612	**11.6640**	1.1086	1.0373	0.9966	0.8417
DAX	1.2417	1.2570	1.1018	0.9748	1.1114	1.3360	1.0875	1.0247	0.9965	0.8122
CAC	0.3371	1.4841	1.0630	0.9650	1.1165	**4.1195**	1.0859	1.0389	0.9942	0.6986

CUADRO 4.30: Comparación de los coeficientes estimados para el modelo de Hentschel [1995], porcentaje de NA=10 %

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
IBM	1.0373	1.2794	1.2205	0.9754	1.3039	1.2033	1.0403	0.9866	0.9926	0.7702
APPLE	0.9046	1.9640	1.5713	0.9355	**0.1205**	**-3.4414**	1.0004	0.9942	0.9824	0.5027
GE	0.0573	1.6350	1.2204	0.9893	1.4134	0.1818	1.0208	0.9749	0.9968	0.5988
TEF	0.2508	1.0150	1.2951	0.9622	**-0.1697**	2.5745	1.0655	1.0578	1.0011	1.0439
SAN	0.9937	1.3523	1.1186	0.9926	1.1403	**1.2128**	1.0762	0.9399	0.9954	0.6951
NOVARTIS	1.5423	1.2998	1.1524	0.9995	1.4221	**1.7623**	1.0402	1.0017	0.9965	0.7702
BMW	1.3461	1.5125	1.2489	0.9838	1.1383	-2.1972	1.0262	0.9914	0.9959	0.6562
EAD	0.9750	1.1246	1.3030	0.9778	0.6605	**-2.5088**	1.0361	0.9994	0.9983	0.8891
NASDAQ	0.7359	1.1180	0.8632	1.0153	**2.4411**	**-1.0147**	1.1121	0.9486	0.9969	0.8499
SP500	0.8638	1.2077	0.9497	0.9896	1.2414	0.0345	1.0818	1.0310	0.9973	0.8531
EUROSTOXX	0.1258	1.2745	0.9405	0.9768	1.3456	2.9225	1.1198	1.0660	0.9970	0.8344
IBEX	**1.4478**	1.2150	1.0846	0.9974	1.2117	**9.7260**	1.0702	0.9735	0.9966	0.8010
SMI	**0.8810**	1.2043	0.9631	0.9786	1.2572	**3.1561**	1.0864	1.0358	0.9970	0.8598
DAX	0.9257	1.3553	1.1570	0.9639	0.9277	**2.3442**	1.1032	1.0649	0.9960	0.7847
CAC	0.2774	1.4692	0.9837	0.9716	1.4504	**2.0092**	1.0883	1.0529	0.9945	0.7146

CUADRO 4.31: Comparación de los coeficientes estimados para el modelo de Hentschel [1995], porcentaje de NA=15 %

	μ	ω	α	β	τ_1	τ_2	ψ	σ_y	ϕ	$h2l$
IBM	1.4033	1.7381	1.4836	0.9565	1.7494	6.3984	1.0915	1.0010	0.9863	0.5723
APPLE	1.3099	2.1497	1.6300	0.9324	**0.1233**	**-4.1010**	1.0176	0.9384	0.9801	0.4368
GE	0.0104	1.9114	1.3361	0.9786	1.2760	-0.5222	1.0518	1.0103	0.9962	0.5284
TEF	0.4520	1.2708	1.4727	0.9350	**-0.0834**	**-4.9652**	1.0712	1.0970	0.9967	0.8621
SAN	1.6457	1.9700	1.2593	0.9917	1.9407	**9.2791**	1.1075	0.9324	0.9921	0.4733
NOVARTIS	**1.1719**	1.3674	1.0549	0.9958	1.3011	0.8432	1.0649	0.9729	0.9956	0.7106
BMW	**0.6079**	1.8334	1.2936	0.9803	**2.0575**	-0.6446	1.0125	0.9985	0.9946	0.5477
EAD	1.0811	1.4004	1.1745	0.9832	**4.6219**	0.6668	1.0302	0.9473	0.9951	0.6783
NASDAQ	1.1275	1.4882	1.0996	0.9998	2.8384	**-8.0831**	1.1238	0.9542	0.9925	0.6374
SP500	0.8488	1.8010	0.9260	0.9713	1.2588	**0.0098**	1.1312	1.1547	0.9933	0.6356
EUROSTOXX	1.5072	1.2038	0.8353	0.9853	2.2035	**1.8255**	1.1643	1.0114	0.9971	0.8401
IBEX	**0.8216**	1.5911	1.0721	0.9957	1.8174	-1.2266	1.1262	0.9366	0.9929	0.5824
SMI	**0.6383**	1.4525	0.8908	0.9765	2.6305	**3.1829**	1.1825	1.0395	0.9939	0.7156
DAX	0.7322	1.1832	0.8994	0.9760	**1.8514**	**2.2512**	1.0312	1.0598	0.9981	0.8989
CAC	**0.1736**	2.2695	1.0472	0.9525	1.8182	**-3.2511**	1.1749	1.1749	0.9906	0.5157

CUADRO 4.32: Comparación de los coeficientes estimados para el modelo de Hentschel [1995], porcentaje de NA=30 %

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
IBM	1.3120	1.0526	0.9262	1.0324	0.9992	1.0464	1.0301	1.1361	0.9992	0.9241
APPLE	1.1478	1.3285	0.8633	1.1181	0.9987	1.4321	1.0507	1.5498	0.9987	0.7662
GE	1.0765	1.0522	1.0045	1.0964	0.9999	1.0623	1.0464	1.3871	0.9999	0.9573
TEF	**1.0013**	1.0714	1.1145	1.0067	0.9995	0.6490	1.0160	1.1878	0.9995	0.8708
SAN	1.1931	1.1007	1.2661	0.9919	0.9997	1.0970	1.0888	1.4905	0.9997	0.7908
NOVARTIS	0.8374	1.0246	0.9462	0.9737	0.9994	1.0366	1.0131	1.0877	0.9994	0.9319
BMW	1.1055	1.1073	0.9317	1.0280	0.9994	1.0427	1.0396	1.3627	0.9994	0.7407
EAD	0.9126	1.3833	1.0439	0.8508	0.9972	1.5023	1.0454	1.5456	0.9972	0.6619
NASDAQ	1.1139	0.9558	0.9222	1.0134	1.0004	1.0588	1.0413	1.1353	1.0004	1.0349
SP500	1.0699	1.3310	0.0341	**1.2713**	0.9973	1.3254	1.0127	1.0565	0.9973	0.7203
EUROSTOXX	1.1130	**0.9291**	1.1685	0.9926	1.0000	0.9802	1.0507	1.1748	1.0000	1.0192
IBEX	1.2271	0.9667	1.1863	1.0207	0.9999	1.0132	1.0355	1.1732	0.9999	0.9530
SMI	1.0177	1.4025	1.1417	0.9831	0.9984	1.4009	1.0701	1.0218	0.9984	0.7067
DAX	1.1735	**1.0002**	1.4775	1.0963	0.9997	0.9420	1.0495	1.2128	0.9997	0.8817
CAC	1.1360	**1.0207**	1.2682	**1.0181**	0.9999	**1.1492**	1.0791	1.1737	0.9999	0.9608

CUADRO 4.33: Comparación de los coeficientes estimados para el modelo Lee & Engle [1999], porcentaje de NA=10%

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
IBM	0.7965	1.0442	1.2867	1.0281	0.9995	1.0087	1.0228	1.1426	0.9995	0.9543
APPLE	1.0715	0.9660	0.8655	1.2034	1.0002	1.2616	1.0424	1.5451	1.0002	1.0395
GE	1.1935	0.8373	0.8984	1.0667	1.0006	0.9903	1.0366	1.4004	1.0006	1.2532
TEF	1.3137	0.8698	1.2139	0.9961	1.0004	0.7390	1.0684	1.1951	1.0004	1.0946
SAN	**2.5717**	1.0417	1.1475	0.9786	0.9999	1.2372	1.0775	1.5436	0.9999	0.9530
NOVARTIS	0.9699	1.1355	1.1946	1.1828	0.9983	1.0664	1.0294	1.0816	0.9983	0.8220
BMW	1.1334	0.9016	**1.7699**	**1.8411**	0.9990	0.7189	1.0347	1.2263	0.9990	0.5905
EAD	1.5271	2.0388	0.8355	0.9820	0.9958	1.5457	0.9498	1.5671	0.9958	0.4820
NASDAQ	1.3047	1.3921	0.3965	1.0586	0.9962	1.5103	1.1109	1.1189	0.9962	0.6746
SP500	1.1013	1.8514	0.0141	**1.4646**	0.9950	2.2720	1.1151	1.0485	0.9950	0.4817
EUROSTOXX	1.2349	0.8080	1.4556	1.0568	1.0000	0.7809	1.0587	1.1311	1.0000	0.9908
IBEX	1.0804	1.0182	1.3268	1.0119	0.9998	1.0045	1.0772	1.1777	0.9998	0.9095
SMI	1.1219	1.2805	1.6310	0.9567	0.9986	1.2031	1.0885	1.0094	0.9986	0.7531
DAX	1.1933	**0.9127**	1.2889	1.0840	**1.0003**	1.0285	1.0537	1.3069	1.0003	1.1293
CAC	1.3008	0.9513	1.5149	1.0496	1.0001	1.1039	1.0814	1.1849	1.0001	1.0389

CUADRO 4.34: Comparación de los coeficientes estimados para el modelo Lee & Engle [1999], porcentaje de NA=15%

	μ	ω	α	β	φ	κ	ψ	σ_y	ϕ	$h2l$
IBM	1.2096	1.3455	1.2282	0.9928	0.9974	1.5803	1.0594	1.1452	0.9974	0.7457
APPLE	0.9597	0.8207	1.4771	0.8922	1.0018	1.4058	1.0478	1.5741	1.0018	1.3238
GE	1.1651	0.9136	**1.4276**	**1.6511**	1.0009	1.0986	1.0692	1.4740	1.0009	1.4075
TEF	1.7903	0.3939	**2.2782**	1.0975	1.0063	0.3929	1.0684	1.2132	1.0063	2.5421
SAN	**2.3193**	1.3162	1.6075	0.9475	0.9997	1.3504	1.1006	1.5587	0.9997	0.7834
NOVARTIS	1.1530	0.7737	**2.1295**	0.8977	1.0028	0.8508	0.9465	1.0996	1.0028	1.2959
BMW	1.1381	1.2651	**2.2498**	**2.3595**	0.9986	0.9227	1.0632	1.2244	0.9986	0.4399
EAD	1.4152	1.2014	1.6009	**1.2726**	0.9986	1.2098	1.0396	1.5570	0.9986	0.8279
NASDAQ	1.1945	1.3166	1.6139	1.3778	0.9963	**1.2767**	1.1154	1.1073	0.9963	0.6819
SP500	0.9829	1.0375	0.0521	**1.5757**	0.9987	1.3090	1.0912	1.0431	0.9987	0.8635
EUROSTOXX	1.6857	0.9938	**1.4467**	0.9814	0.9995	**0.8899**	1.0755	1.1216	0.9995	0.8114
IBEX	0.9408	0.8877	1.5364	1.0176	0.9997	0.9508	1.1903	1.1168	0.9997	0.8323
SMI	1.4544	1.0056	1.4240	1.0059	0.9994	0.9635	1.1094	0.9748	0.9994	0.8913
DAX	1.1551	**0.6440**	3.2648	1.7282	1.0008	0.7422	1.0321	1.2247	1.0008	1.3064
CAC	**1.0511**	**0.6372**	**2.3346**	**1.4321**	**1.0008**	**0.6871**	1.0767	1.1387	1.0008	1.2934

CUADRO 4.35: Comparación de los coeficientes estimados para el modelo Lee & Engle [1999], porcentaje de NA=30 %

Anexo III: Gráficos

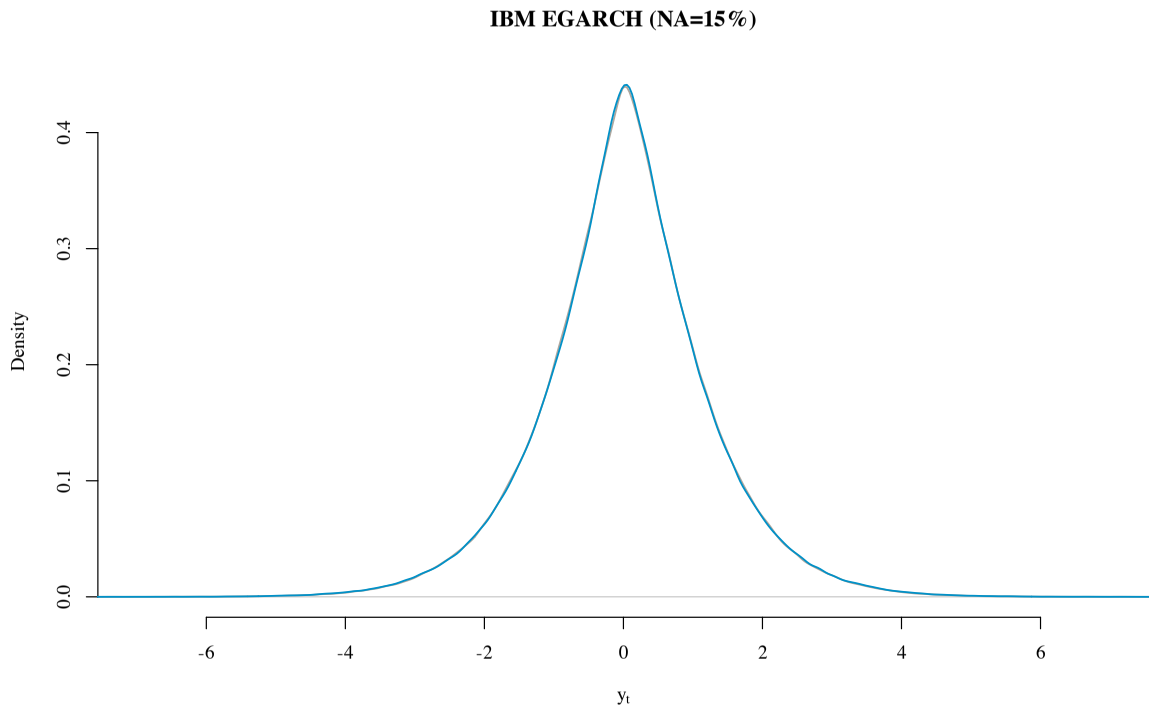

FIGURA 4.1: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para IBM

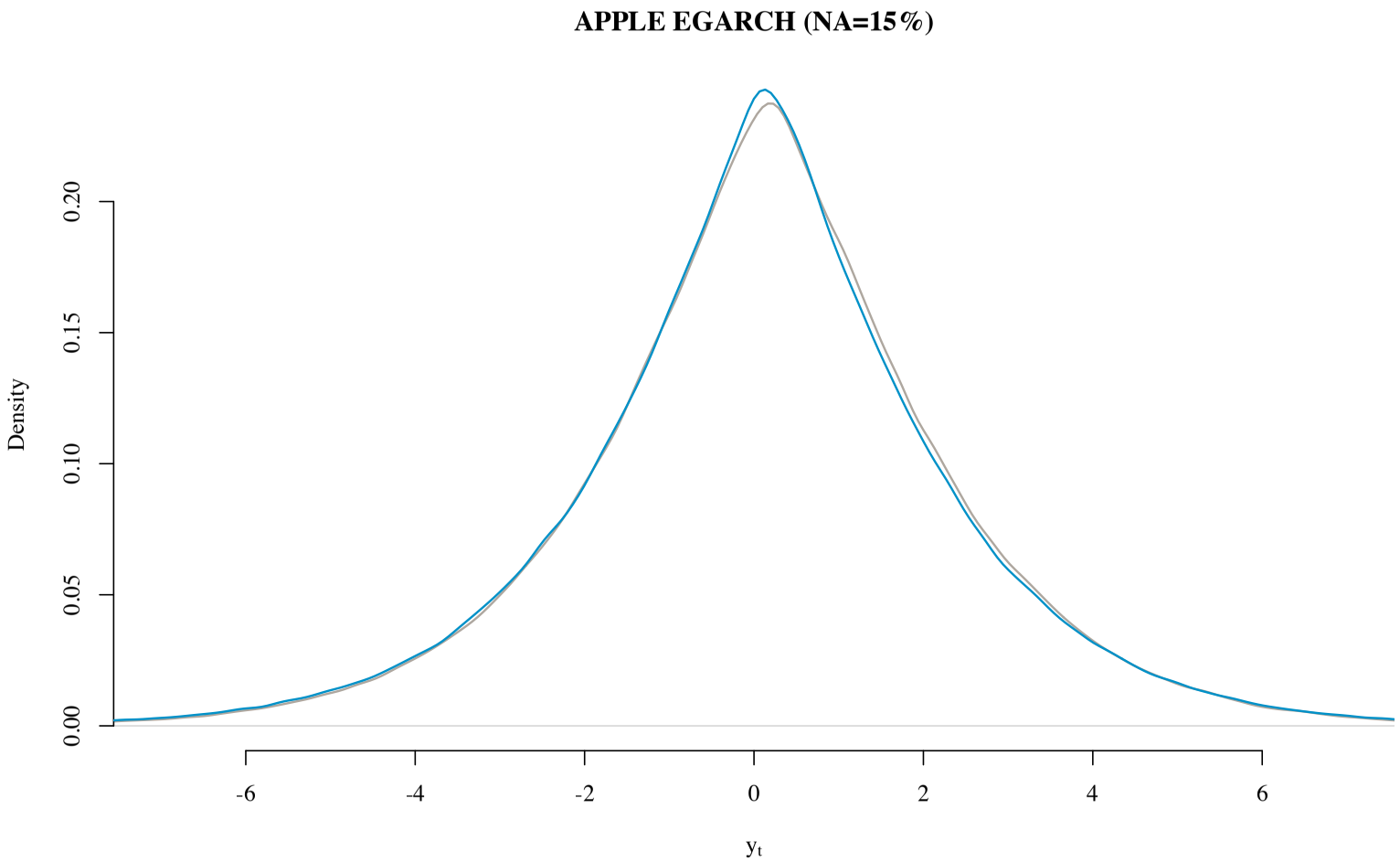

FIGURA 4.2: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para Apple

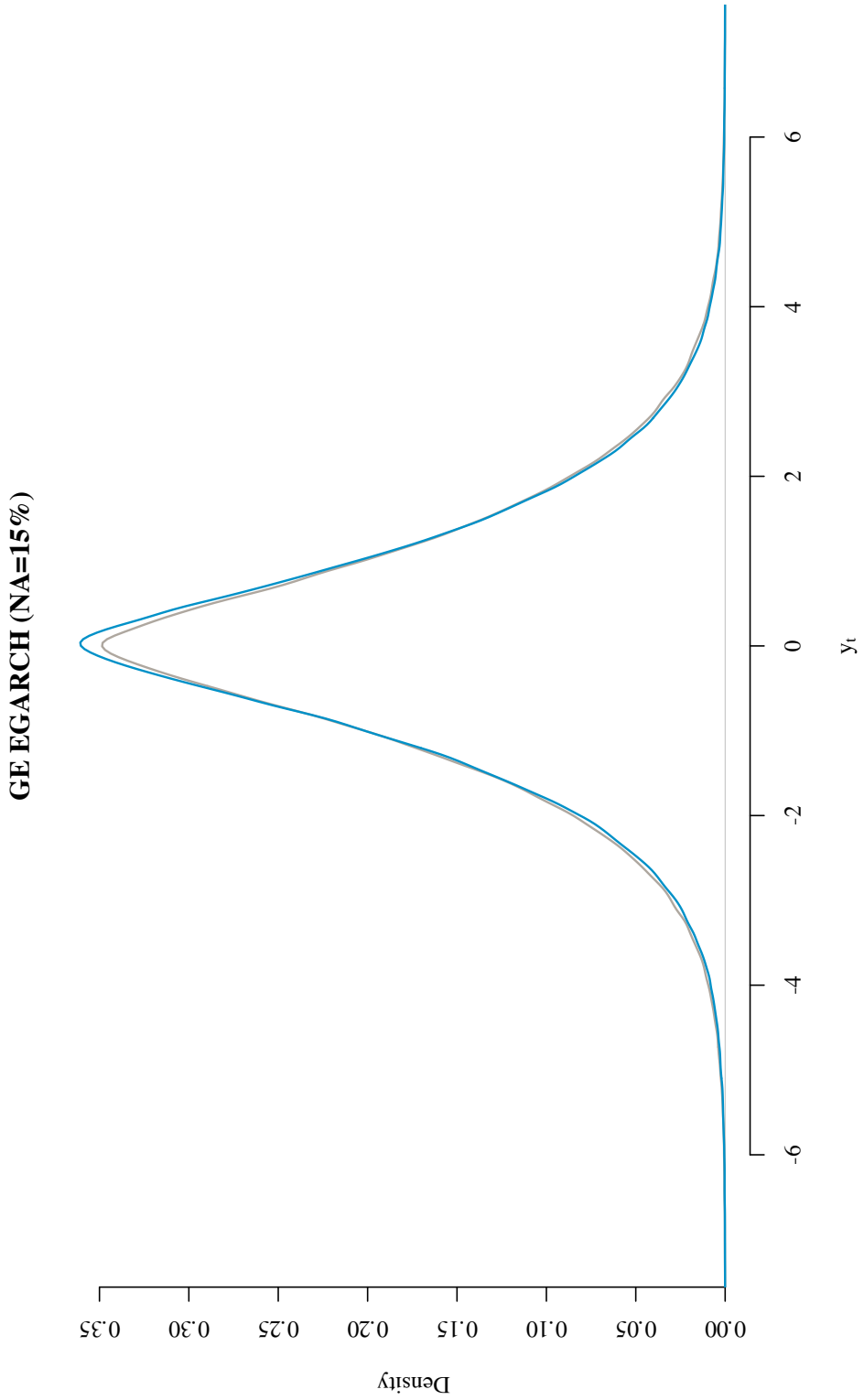

FIGURA 4.3: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para GE

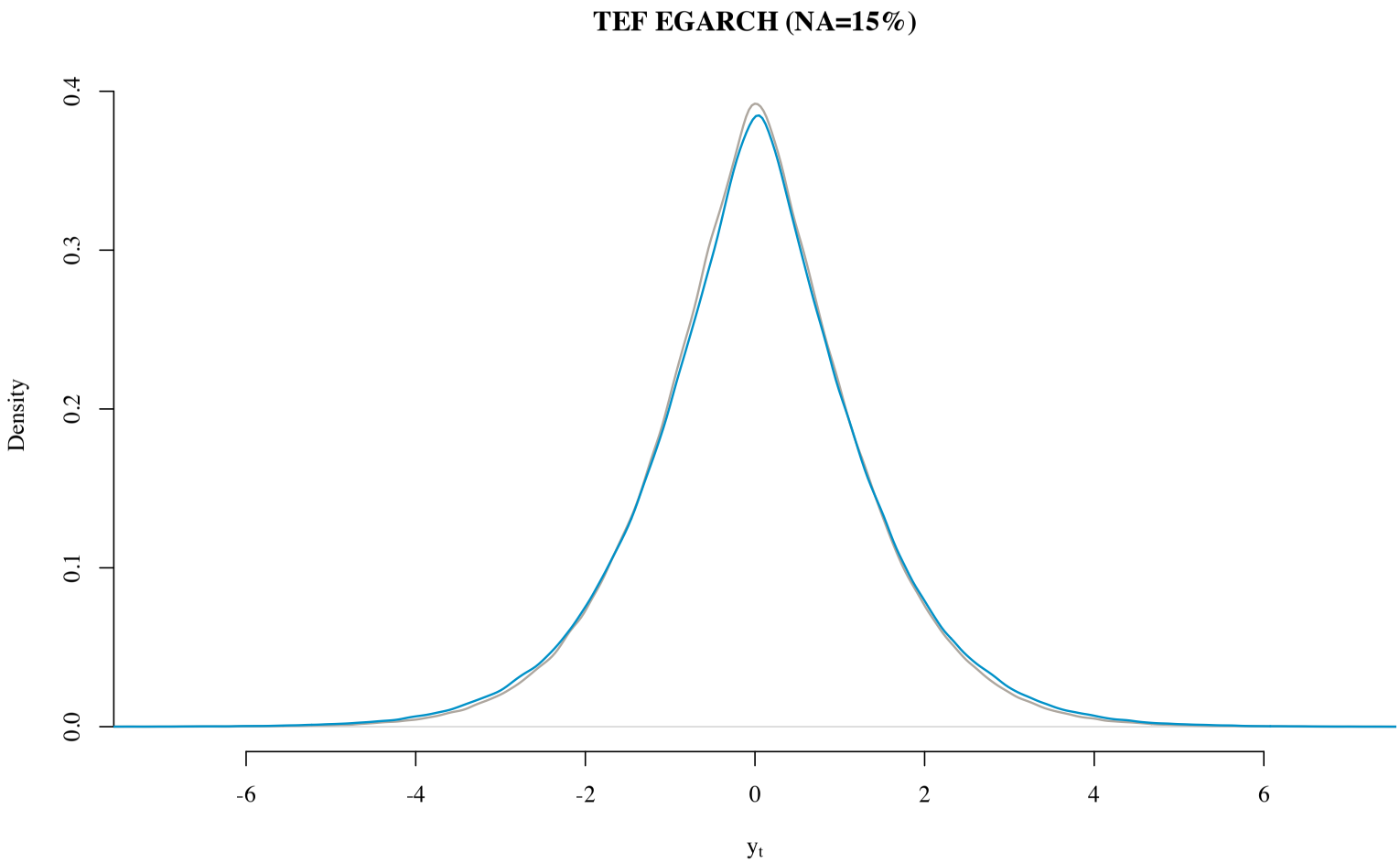

FIGURA 4.4: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para TEF

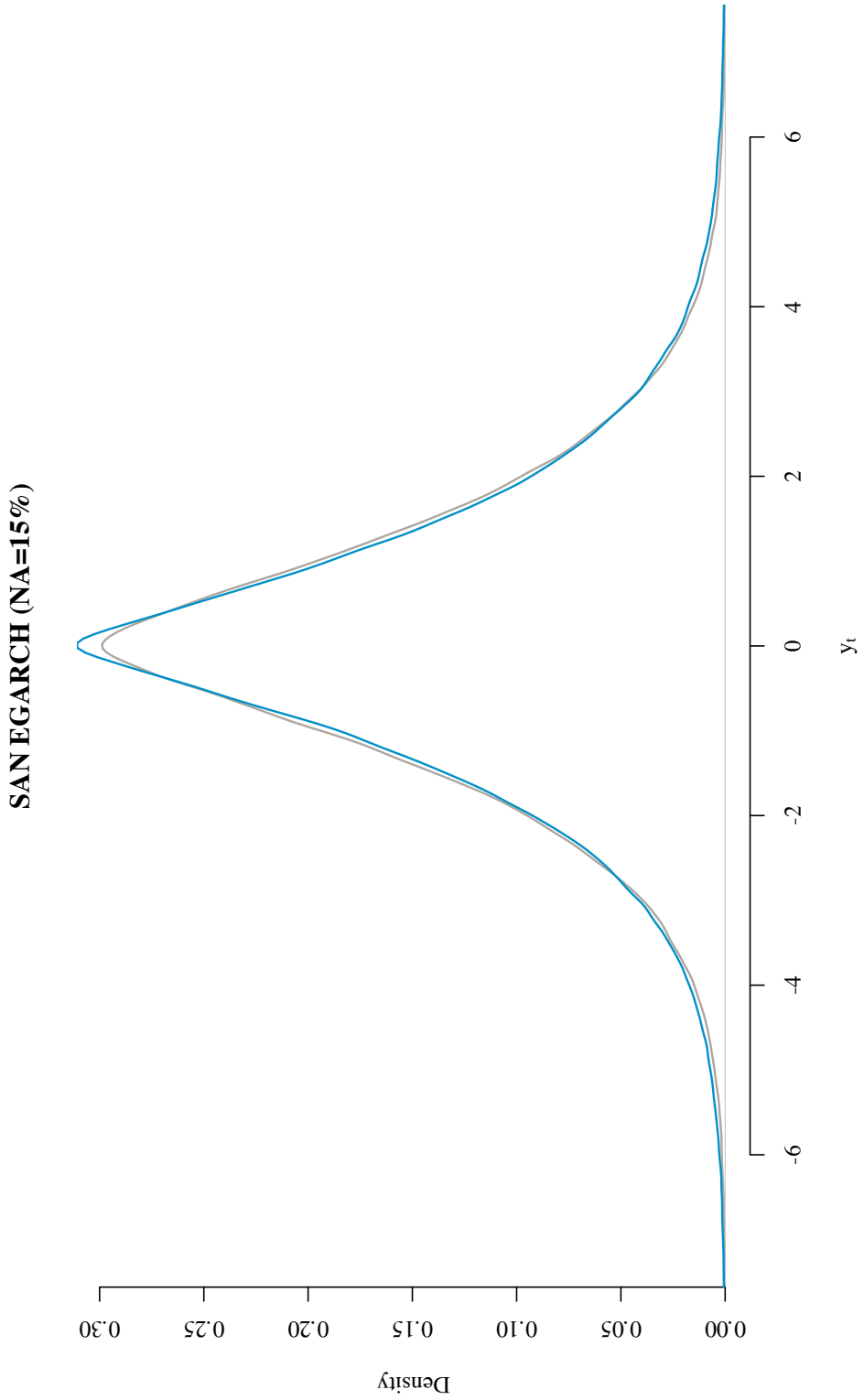

FIGURA 4.5: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para SAN

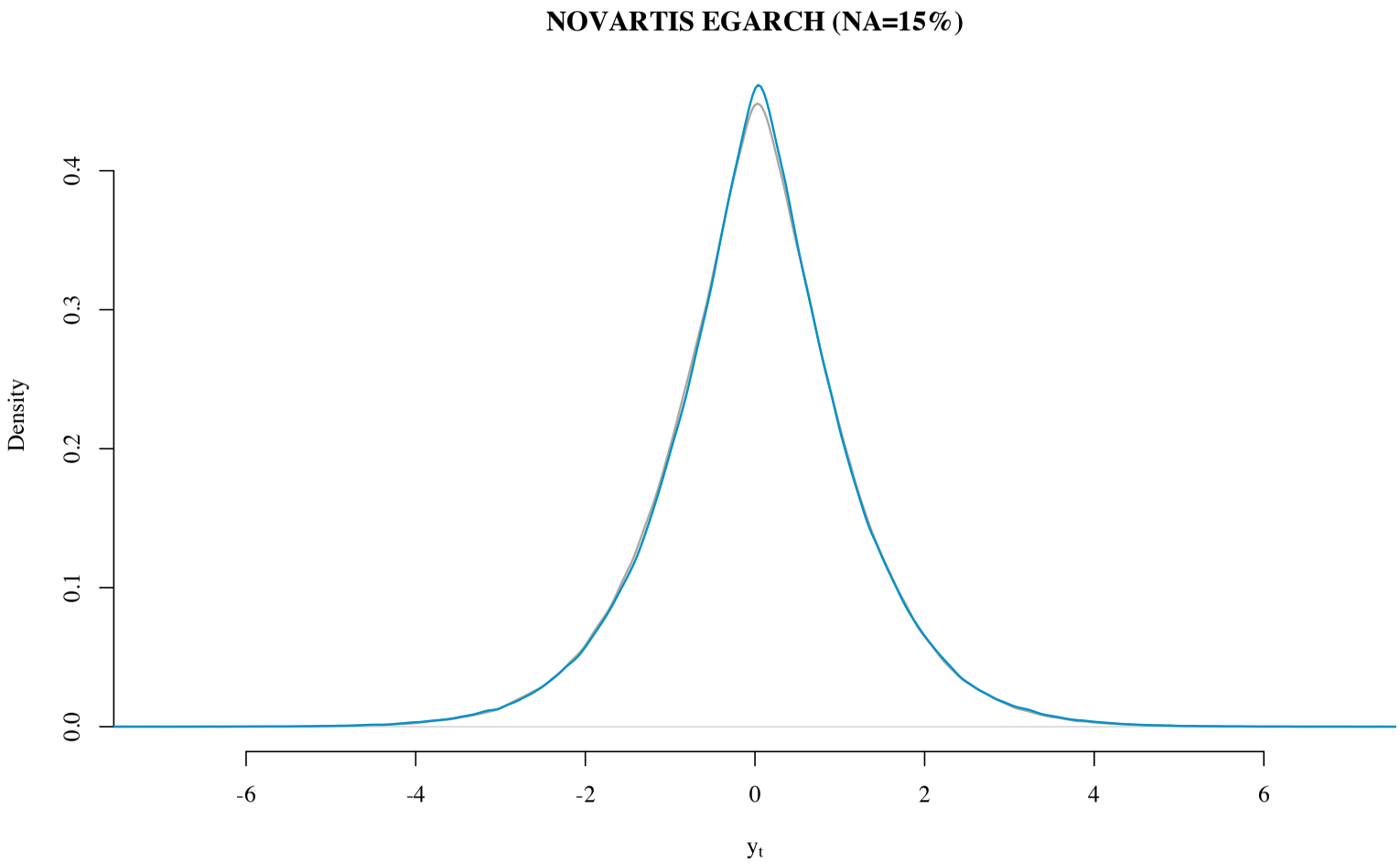

FIGURA 4.6: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para NOVARTIS

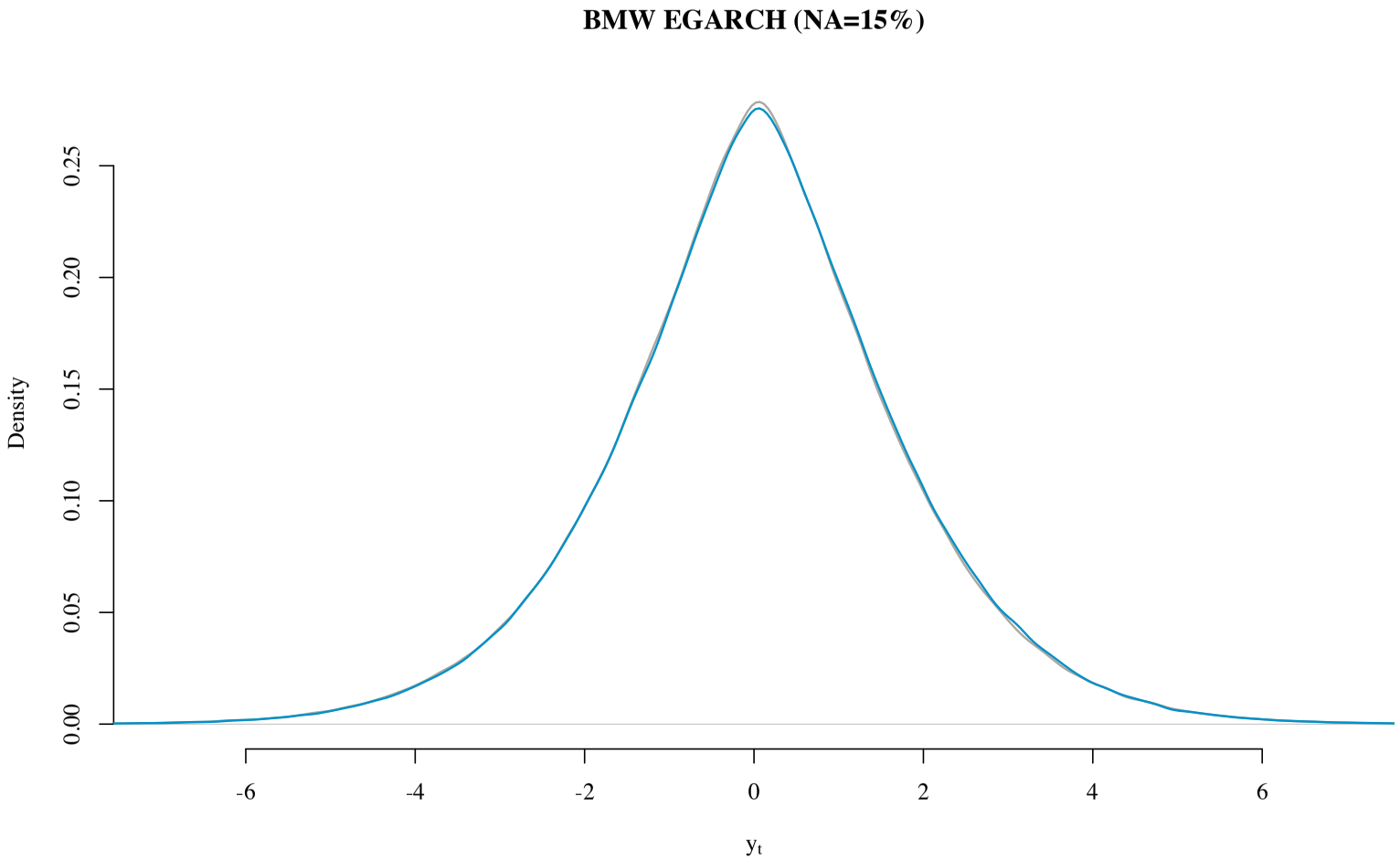

FIGURA 4.7: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para BMW

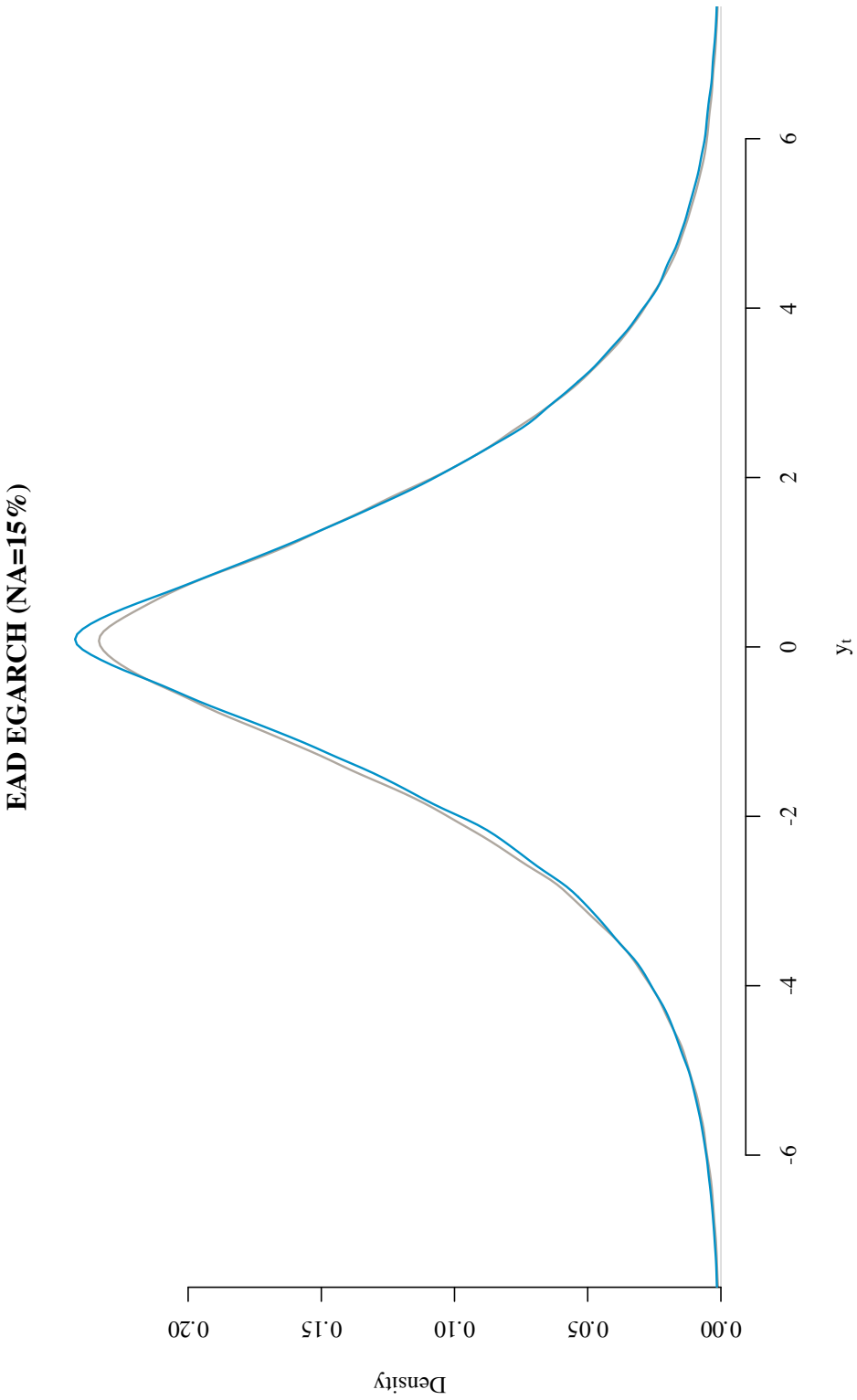

FIGURA 4.8: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para EAD

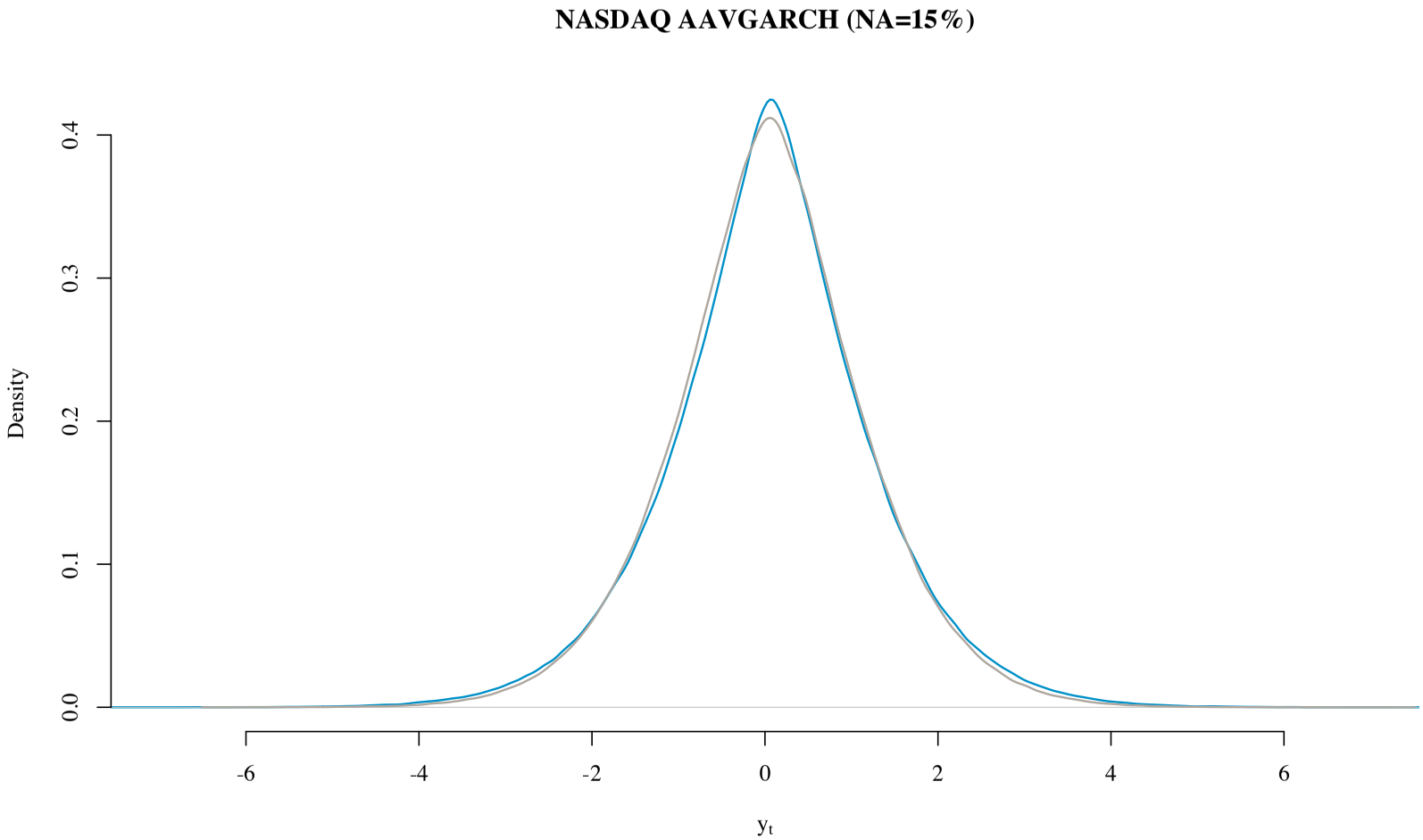

FIGURA 4.9: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para NASDAQ

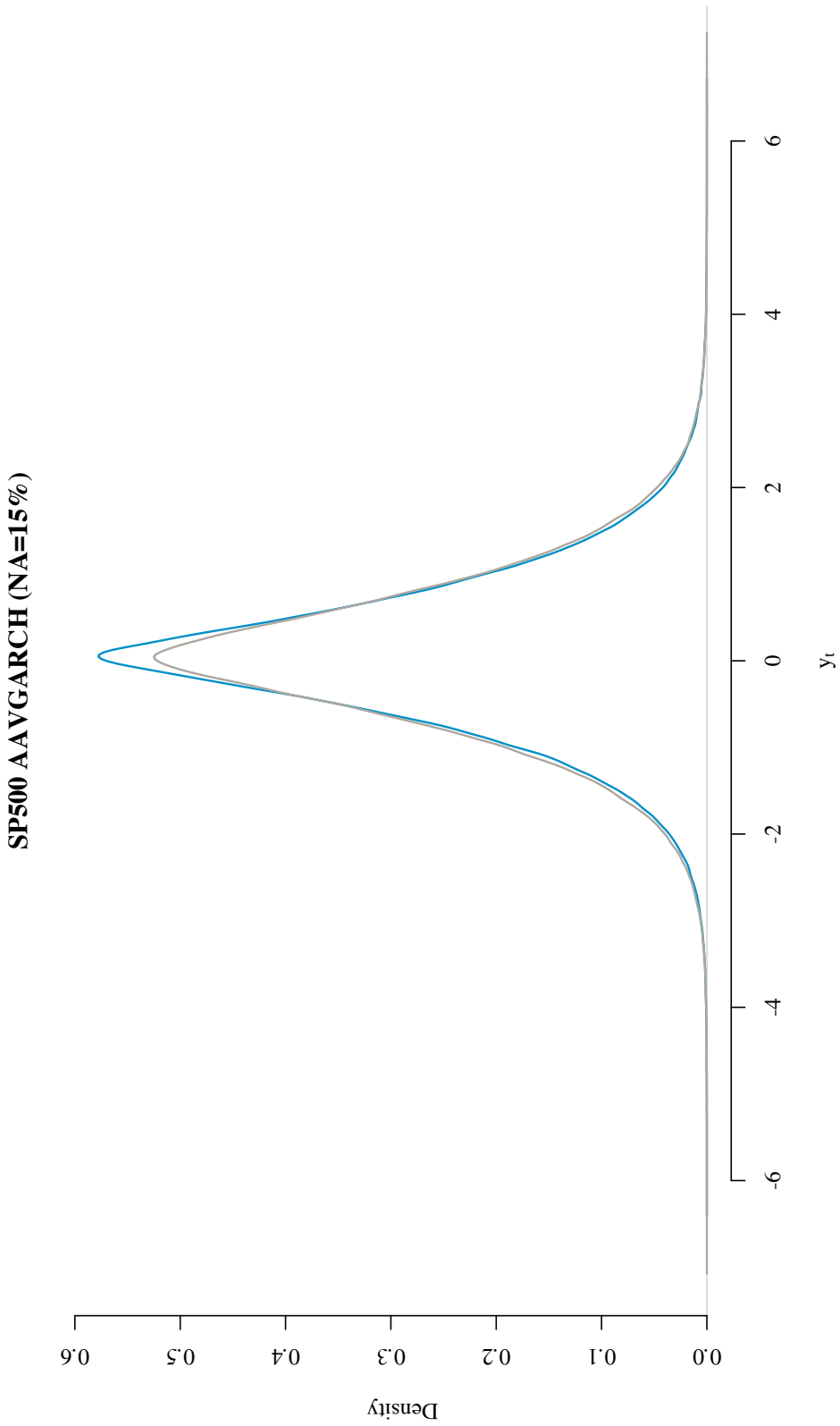

FIGURA 4.10: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para SP500

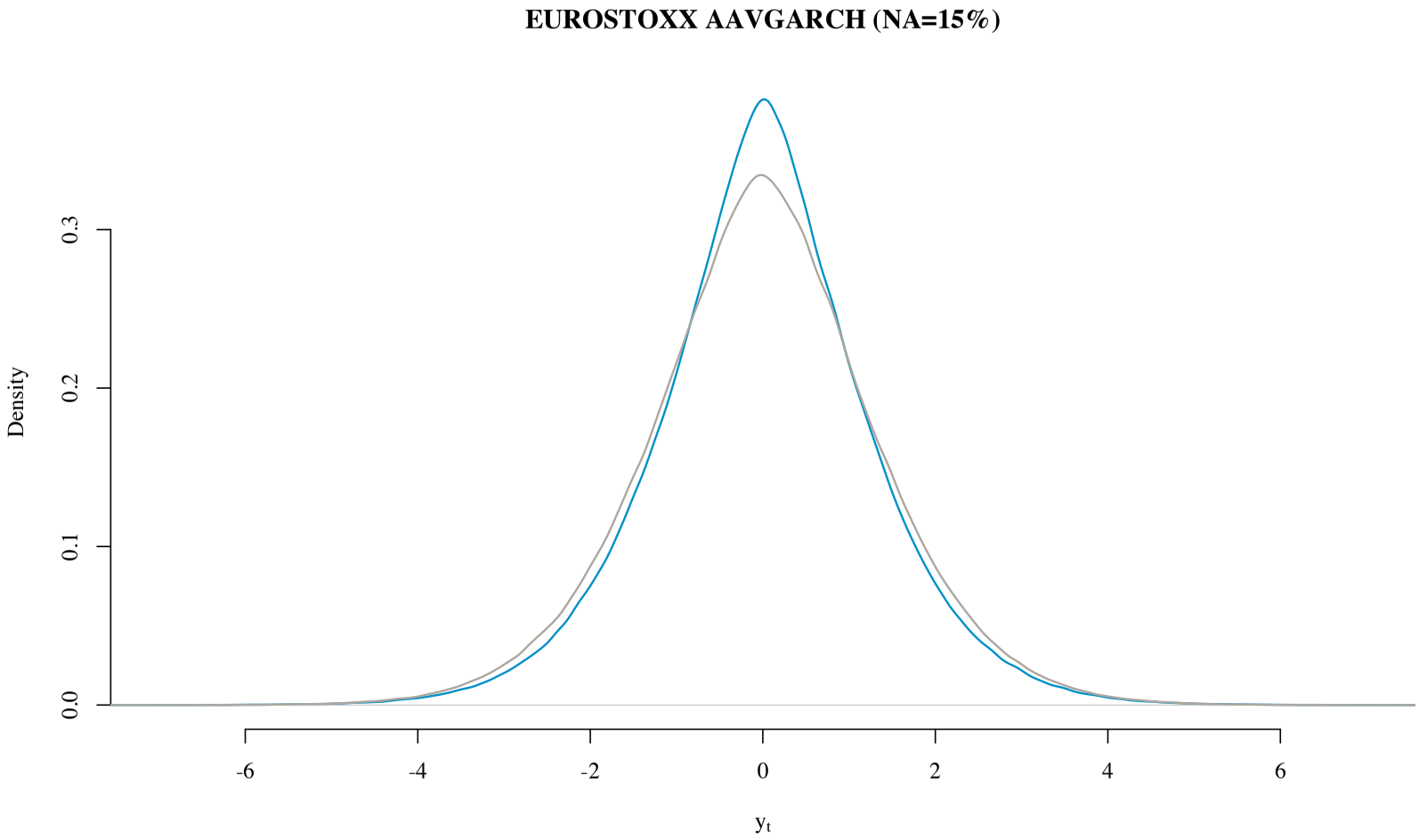

FIGURA 4.11: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para EUROSTOXX

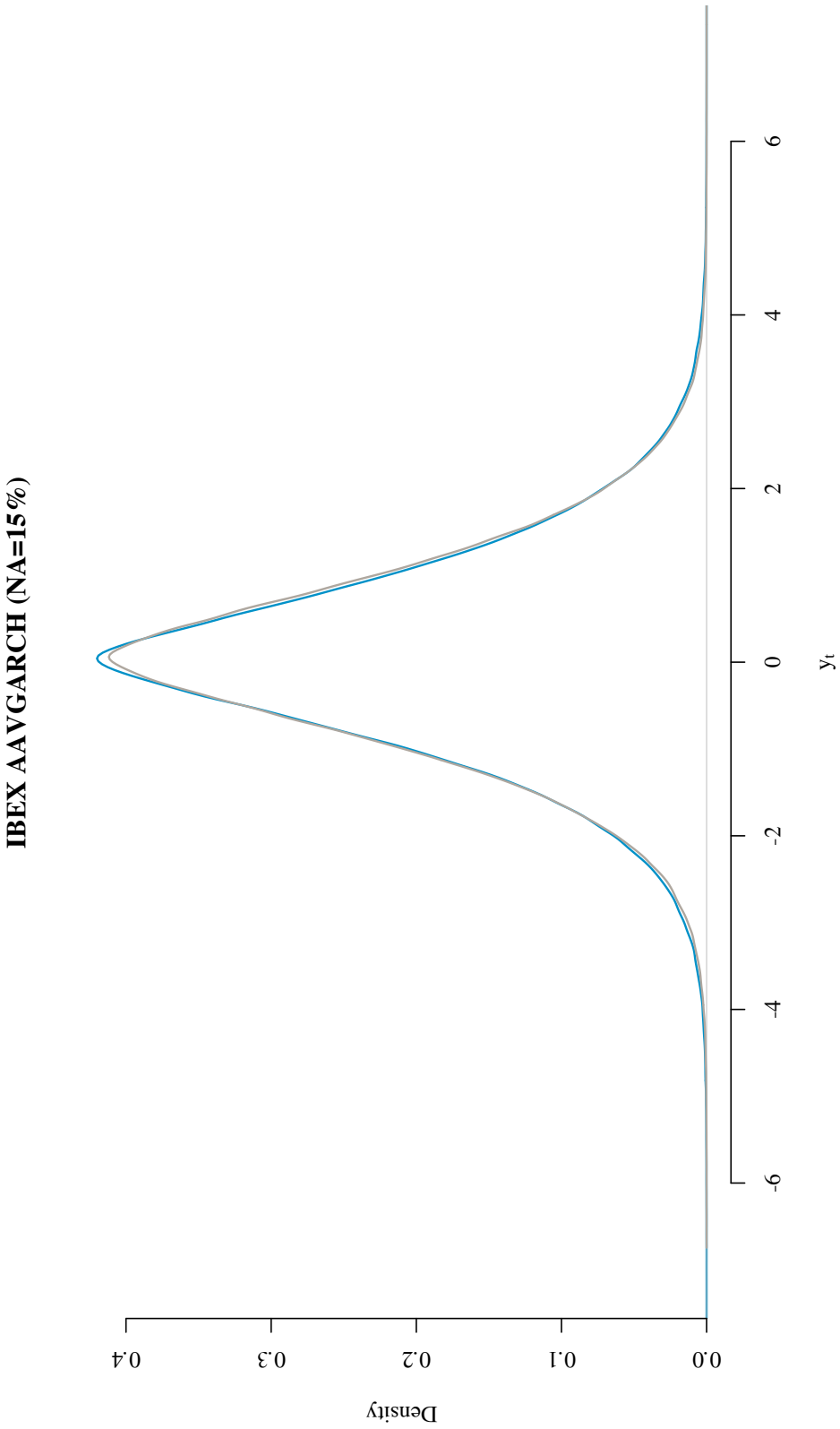

FIGURA 4.12: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para IBEX

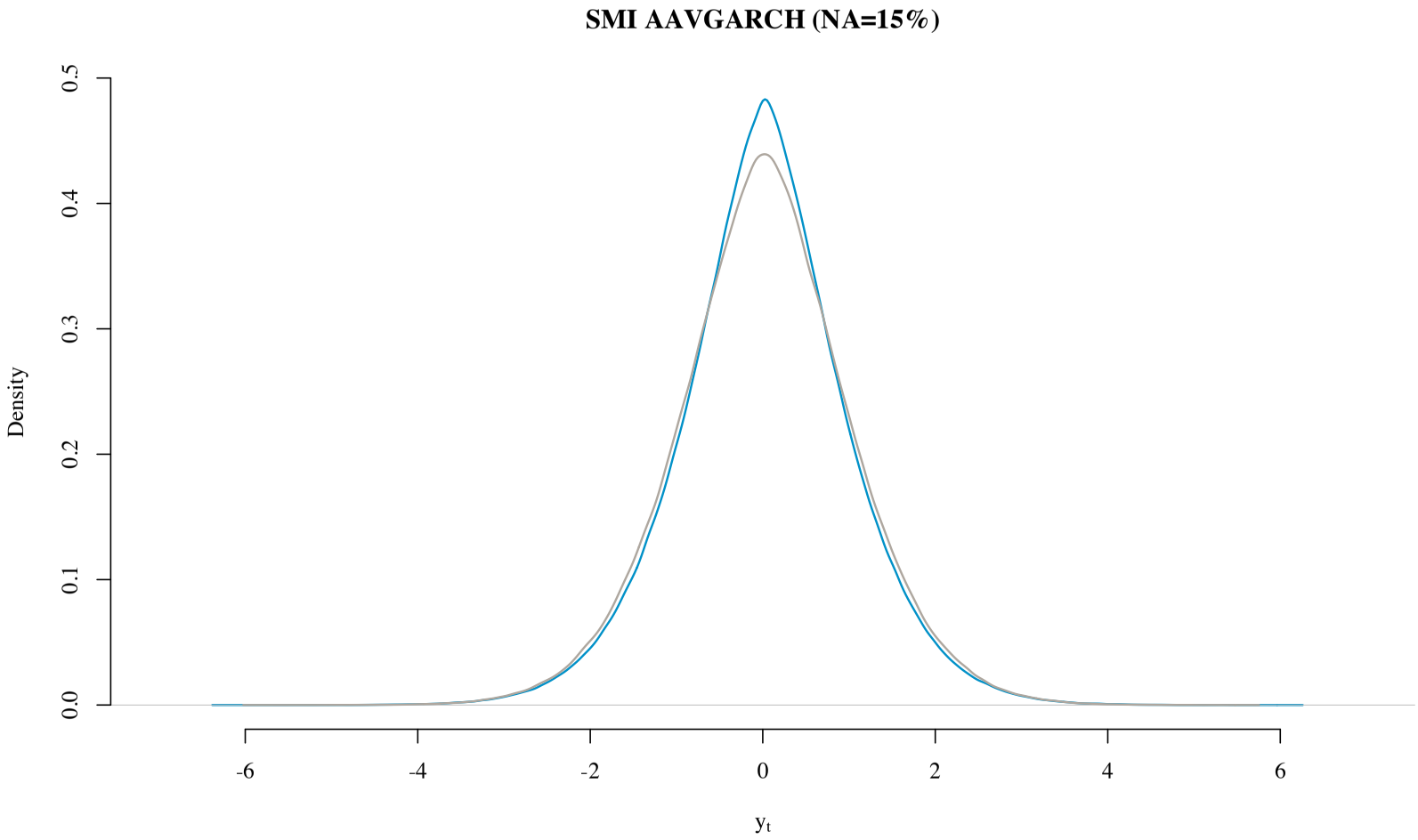

FIGURA 4.13: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para SMI

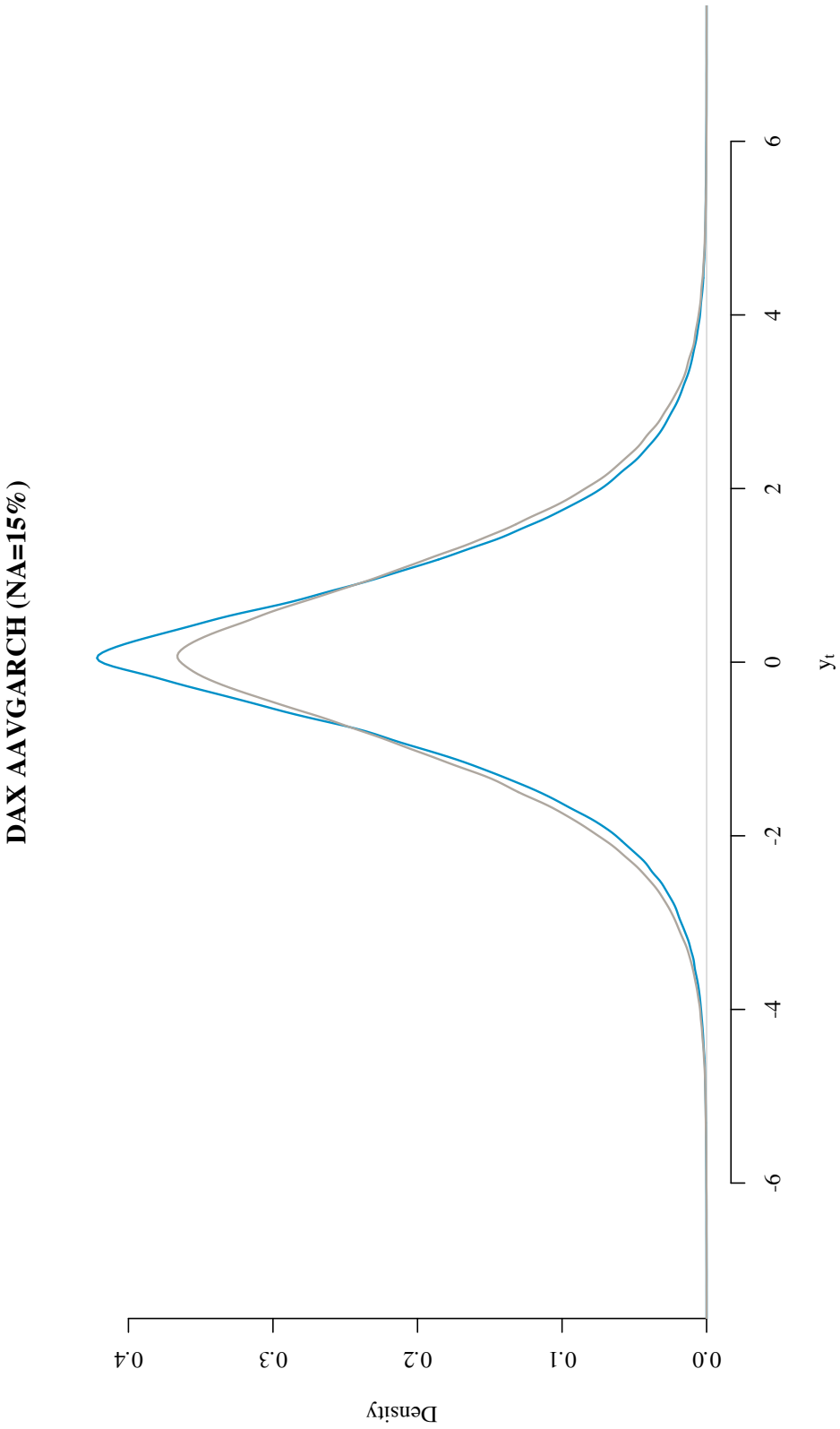

FIGURA 4.14: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para DAX

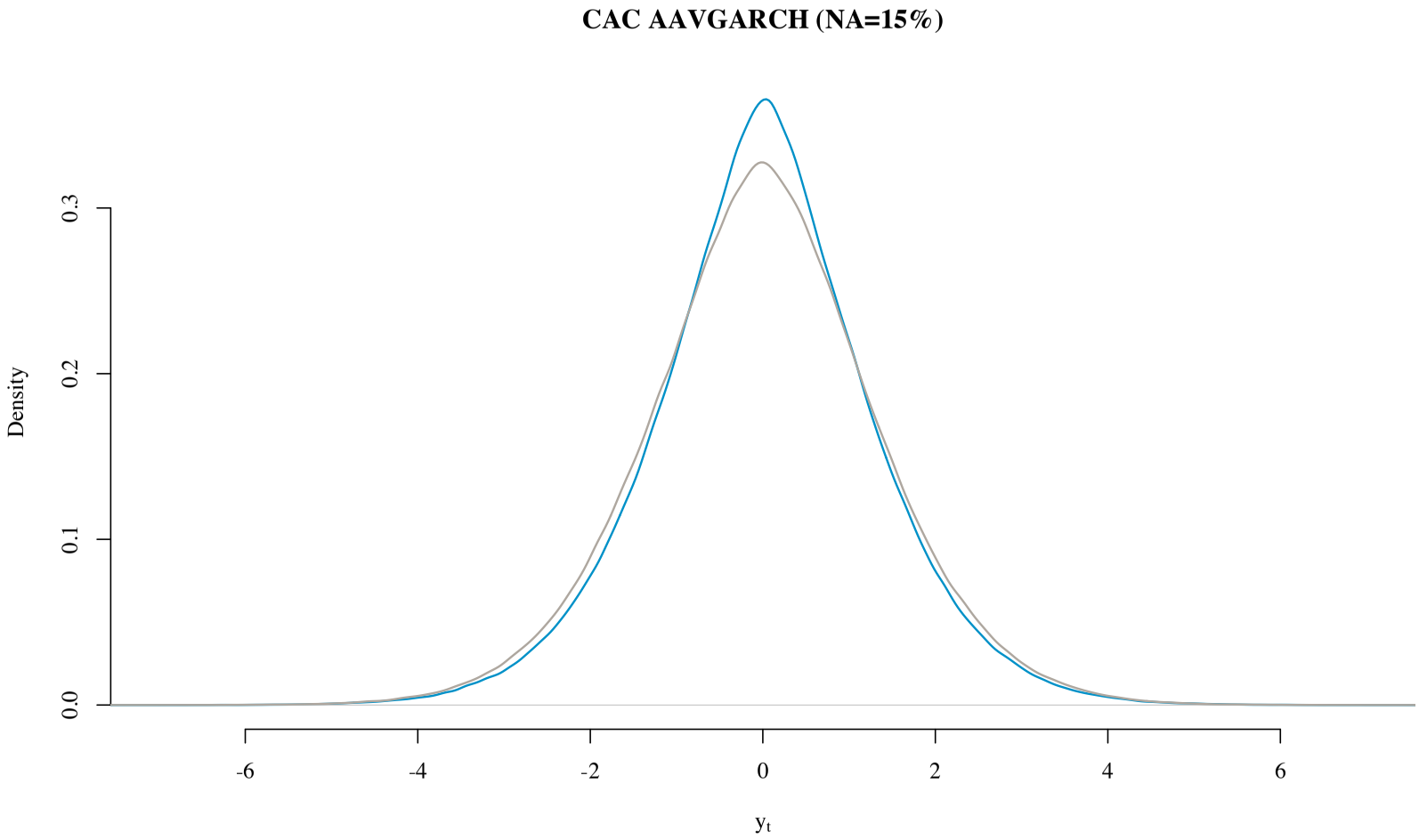

FIGURA 4.15: Comparación de las densidades de los datos (gris para los datos reales, azul para los datos simulados) para CAC

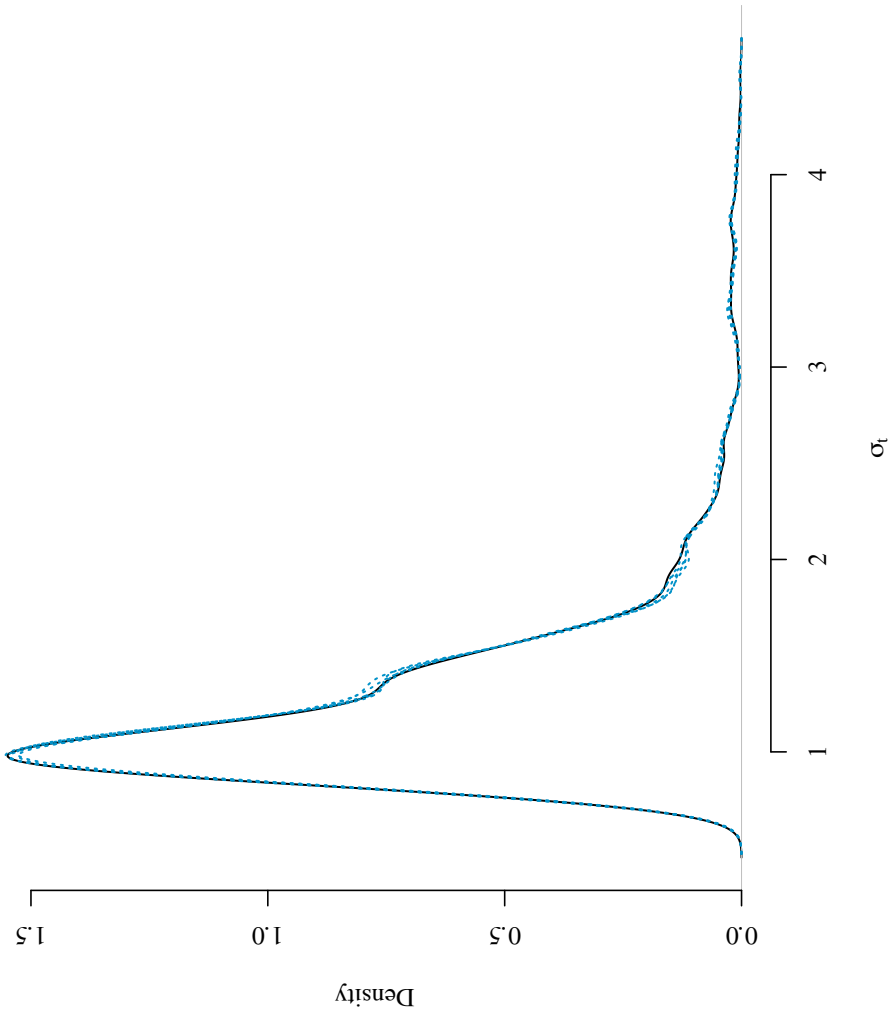

FIGURA 4.16: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para IBM. Porcentaje de NA=15%

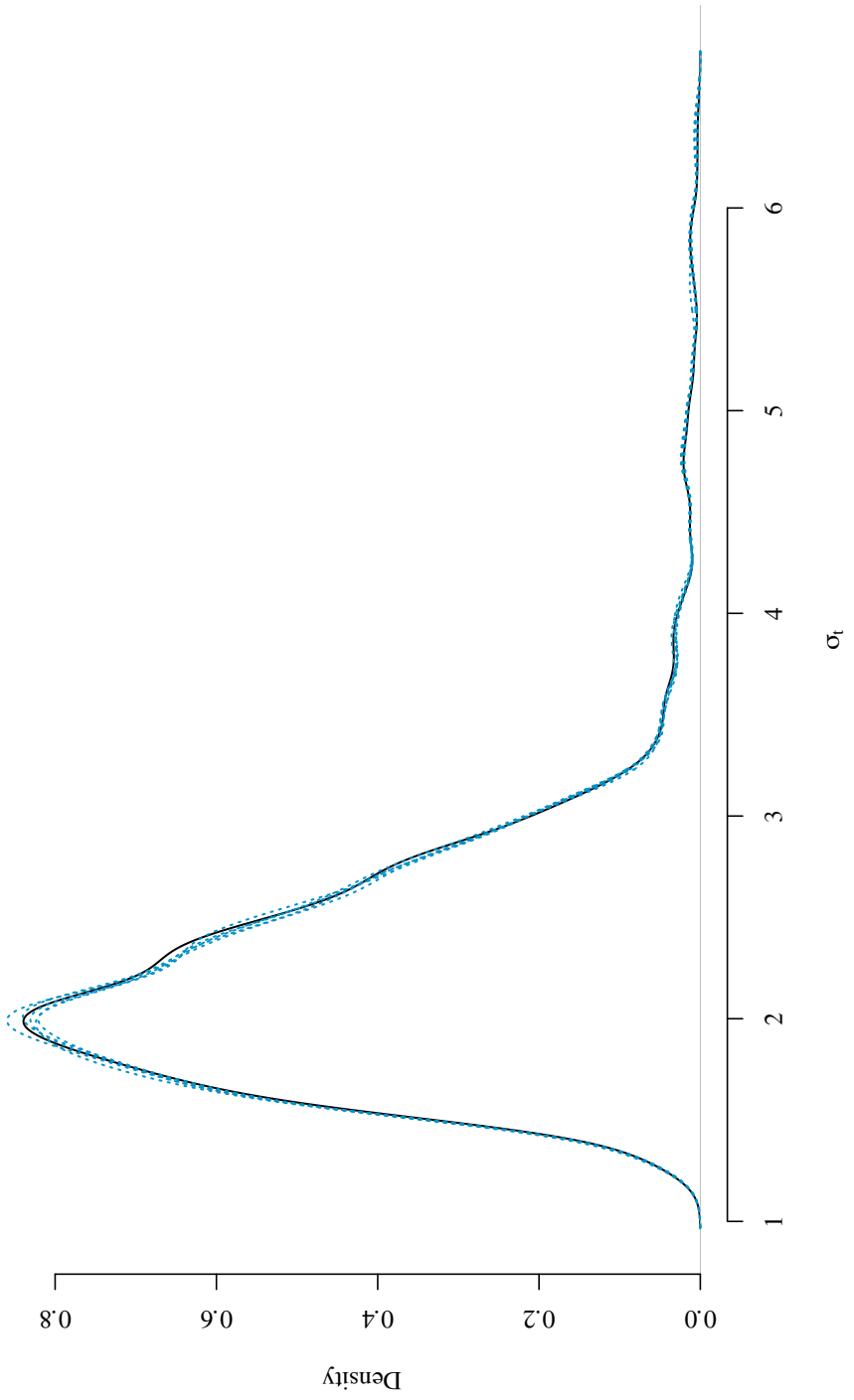

FIGURA 4.17: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para Apple. Porcentaje de NA=15%

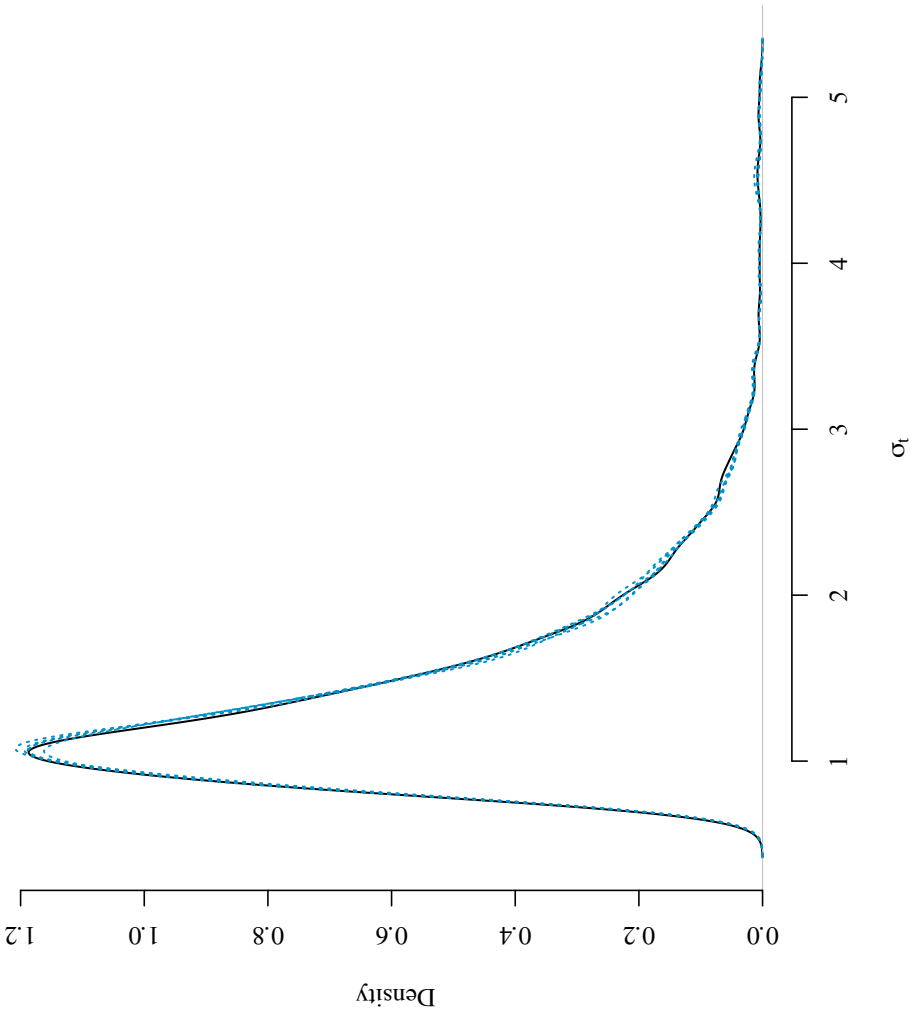

FIGURA 4.18: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para Telefonica. Porcentaje de NA=15%

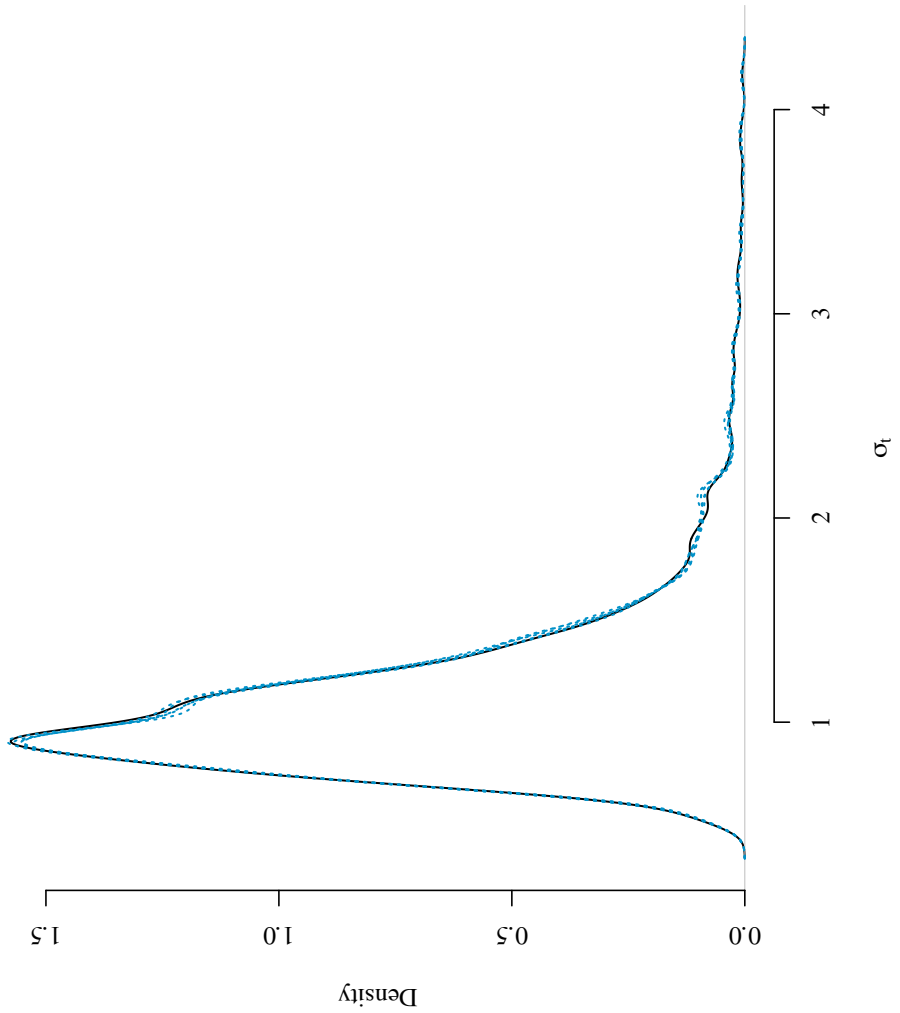

FIGURA 4.19: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para Novartis. Porcentaje de NA=15%

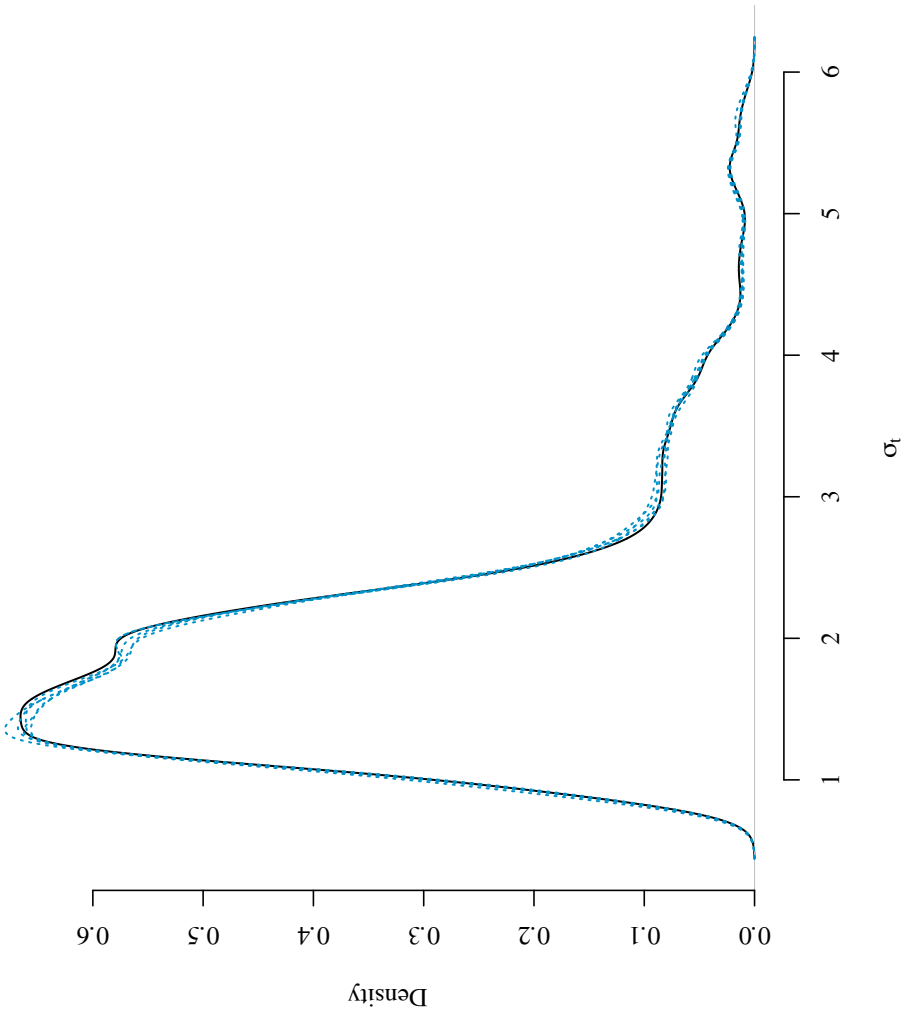

FIGURA 4.20: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para BMW. Porcentaje de NA=15 %

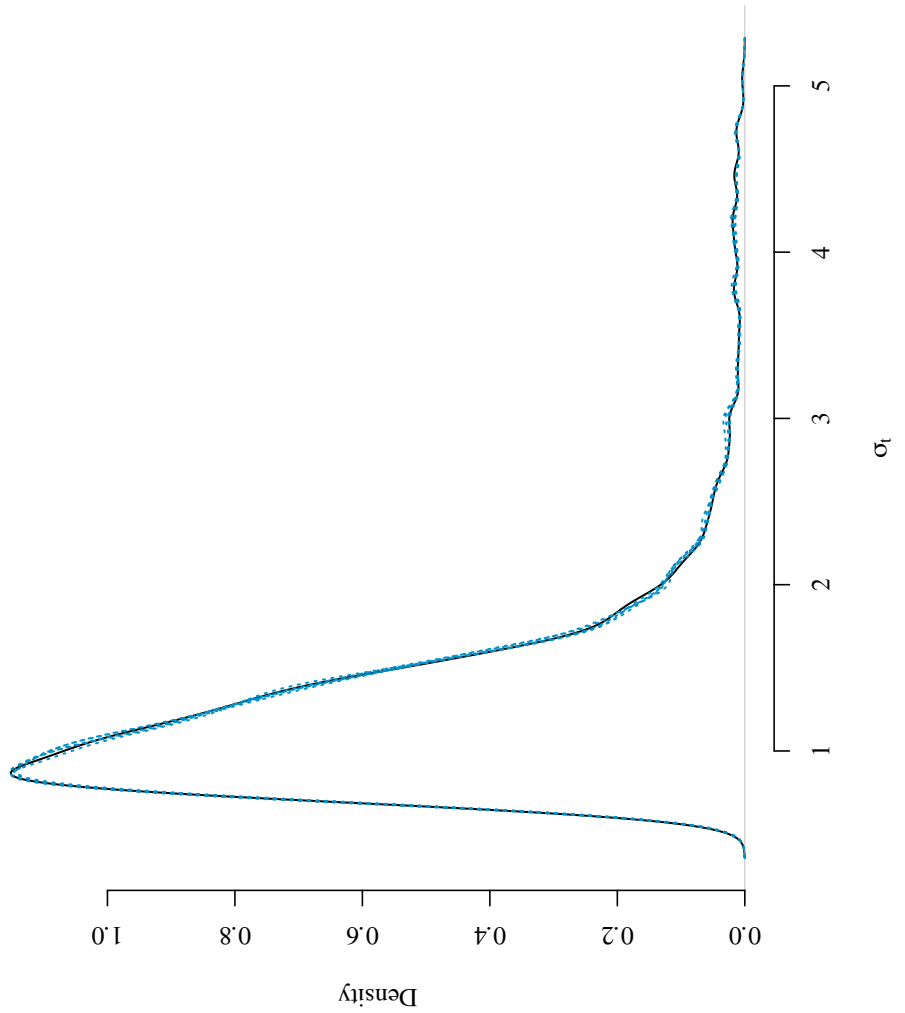

FIGURA 4.21: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para Nasdaq. Porcentaje de NA=15%

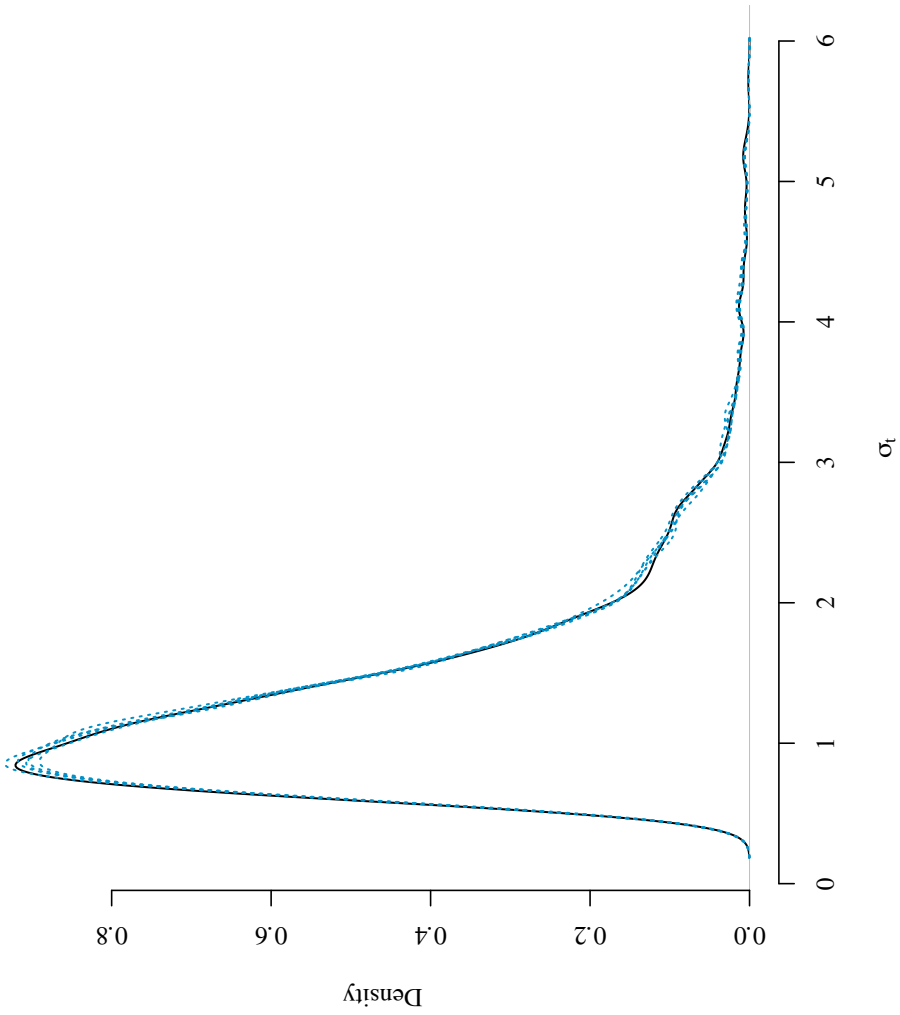

FIGURA 4.22: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para Eurostoxx. Porcentaje de NA=15 %

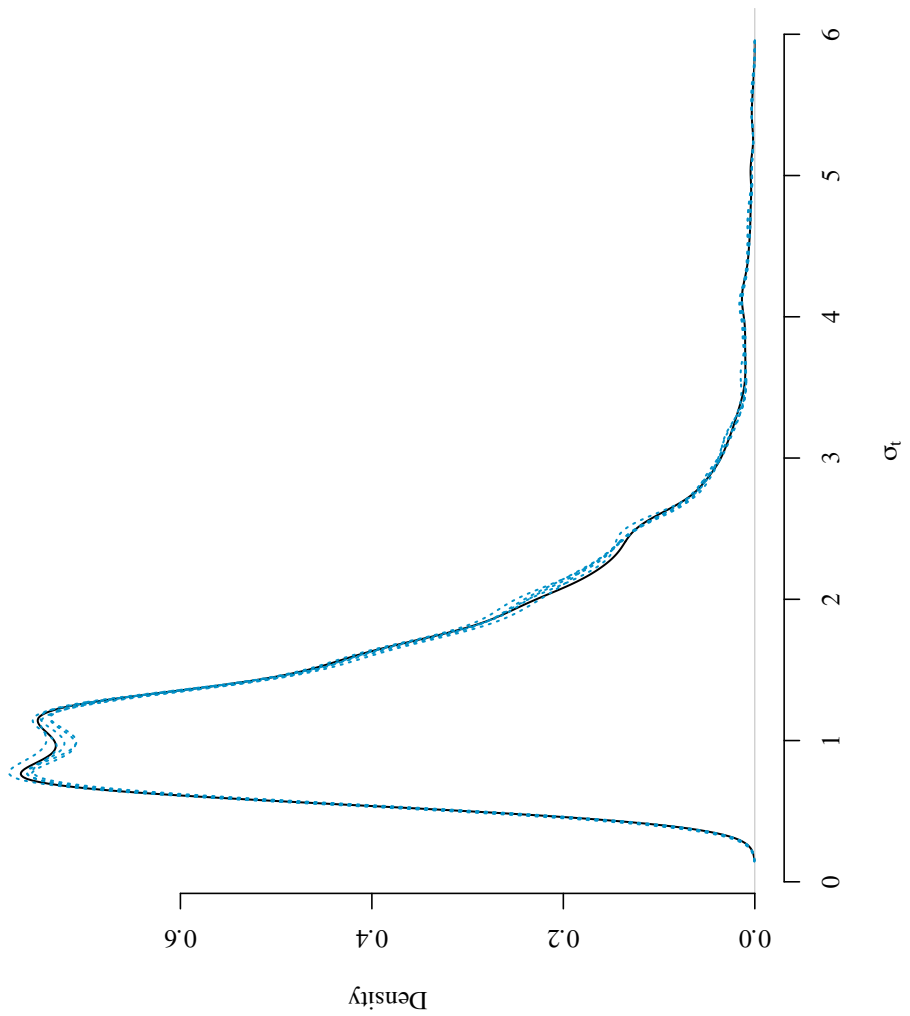

FIGURA 4.23: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para IBEX35. Porcentaje de NA=15%

FIGURA 4.24: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para SMI. Porcentaje de NA=15 %

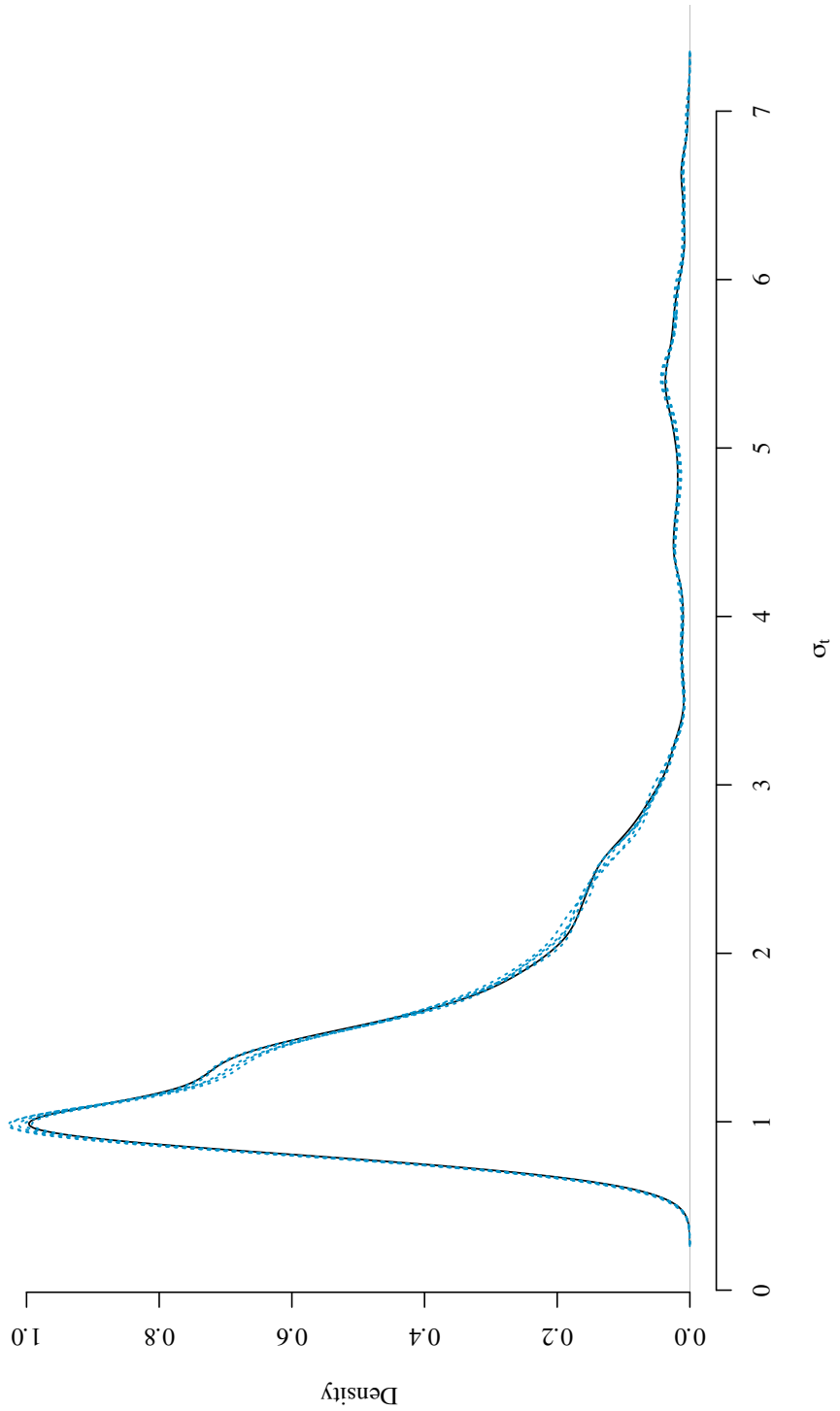

FIGURA 4.25: Comparación de las densidades de σ_t (negro para los datos reales, azul para las imputaciones) para DAX. Porcentaje de NA=15%

Conclusiones

Retrospectivas

I. con respecto a la elección y uso de R

- **R** ha permitido llevar a cabo el trabajo que aquí se presenta gracias a su excelencia en la interacción informática - tema de investigación (definido éste en las pp. 14-18). Creemos que la mentada interacción ha resultado beneficiosa para **R** debido a que ha dado lugar a la construcción de una librería (**`mists`**), que presentaremos al portal CRAN para aspirar a su aceptación como librería contribuida; en dicho portal no existe todavía un librería contribuida de imputación múltiple en sección longitudinal. Con esto pensamos que hemos alcanzado unos de los objetivos que nos habíamos propuesto en la Tesis (objetivo 5).

  También la interacción se ha dejado sentir beneficiosamente sobre el desarrollo del tema de investigación en varios puntos, como se verá seguidamente.

II. con respecto a la Imputación múltiple y el número de imputaciones necesarias para su correcta implementación en sección cruzada (inferencia de Rubin)

- El problema del número mínimo de imputaciones necesario para determinar λ y RE (fracción de información perdida y eficiencia relativa, respectivamente) no está planteado *per se*, debido a que los autores indicaron que utilizaban $m = 5$ y no justificaron el porqué. Nosotros tampoco hemos tratado el problema por vía estrictamente matemática, pero el uso de **R** ha confirmado el número de cinco imputaciones. Basándonos en lo indicado en el epígrafe podemos concluir que $\lambda(m = 2)$, $\lambda(m = 3)$, $\cdots$, $\lambda(m = 5)$ son, en principio, sensiblemente distintos, pero

que, a partir de 5, el valor de λ tiende a estabilizarse en un determinado nivel y que, en consecuencia, RE tiende a 1 cuando $m \to \infty$.

- En Rubin [1987], al tratar la inferencia se habla de λ pero no se discute la "función λ". En tanto que función, λ está relacionado con m a través del siguiente esquema de dependencia que pasamos a describir: λ depende de r (aumento relativo en la varianza debido a la no respuesta) y gl (grados de libertad); a su vez r depende de m y $B/\bar{U}$ (cociente entre las varianzas between y within) y a su vez esta última quedará determinada cuando se fije m. Por su parte, gl depende también de m y r, por lo que teniendo en cuenta lo que acabamos de exponer, λ depende de m.

  Fijado m, quedará unívocamente determinado el cociente $B/\bar{U}$ relativo al dataset. Concluimos pues que λ, para un determinado dataset, depende de, además de m, de:

  1. el modelo de estimación usado.
  2. el tipo de estimación empleado (aquí sólo la puntual)

  Esta conclusión no la hemos sabido ver en Rubin [1987] ni en la literatura posterior, y tampoco estaba formulada como uno de los objetivos de la Tesis, sino que se ha encontrado al estudiar la inferencia de Rubin.

III. con respecto a la Imputación Múltiple y su correcta implementación en STF

Las conclusiones alcanzadas al estudiar la inferencia de Rubin han tenido vocación completadora y, por tanto, aspiran a ser una continuación del trabajo de Rubin; sin embargo, el tránsito al estudio de la imputación múltiple en sección longitudinal hemos debido hacerlo con nuestras solas fuerzas.

- En el epígrafe *3.3* se ponía de manifiesto que al tratar y_t como un todo no se conseguía obtener valores plausibles. Además, la plausibilidad de los valores de y_t exigían que se mantuvieran la **asimetría** y la **leptocurtosis** propias de las STF. A tal efecto utilizamos el Threshold GARCH para escindir y_t en dos procesos distintos: σ_t (volatilidad) e innovaciones ε_t, para posteriormente imputar cada una de ellas mediante algoritmos diferentes (*Approximate Bayesian Bootstrap* y *Gibbs Sampling*). Como puede verse en las figuras del Anexo III, los dos algoritmos aproximan de manera adecuada las funciones de densidad de σ_t e y_t, lo que indicaría que la estrategia empleada es capaz de generar valores plausibles de y_t.

  El análisis de los modelos GARCH propuesto en el capítulo *4* revela que el peso de las colas, ψ, de la distribución condicionada de y_t tras llevar a cabo las imputaciones, se aproxima mucho al coeficiente estimado en caso de completitud de datos. Por consiguiente, estaríamos en disposición de concluir que se ha mantenido la leptocurtosis propia de y_t.

  Dichos modelos GARCH también exhiben comportamiento asimétrico en las innovaciones. Tanto el modelo de Nelson [1991] como el modelo de Hentschel [1995] muestran un comportamiento distinto de la volatilidad ante innovaciones de signo negativo. Sin embargo, destacamos, que en la mayoría de ocasiones, las magnitudes de los coeficientes que capturan la asimetría se reducen. Por lo tanto podemos concluir que la asimetría de la volatilidad se mantiene, pero ésta tiende a suavizarse.

- Los cuadros comparativos del apartado *4.6* indican que ϕ (persistencia del modelo) y σ_y (volatilidad incondicionada) poseen una elevada precisión con respecto a las estimaciones originales. La magnitud de ambos parámetros, por norma general, es superior a la obtenida en el caso de que el dataset sea completo. El aumento de σ_y está en línea con una de las ideas claves de la imputación múltiple de Rubin [1987]: "la varianza de un modelo imputado ha de ser mayor que la resultante de tener un dataset completamente observado".

- El análisis de los modelos elegidos en la sección *4.2* revela que la especificación que mejor ha funcionado con los datasets imputados es el EGARCH de Nelson [1991]

(de amplia difusión en la literatura). De hecho, los gráficos de densidades (Anexo III) de la distribución de y_t son prácticamente idénticos. Este hallazgo nos sugiere nuevamente, apoyándonos en el modelo EGARCH, la plausibilidad de los valores generados.

- Otro resultado interesante que se desprende de las cuadros comparativos del Anexo II es la distorsión en el cálculo de τ_2 (coeficiente que captura la asimetría de las innovaciones pequeñas en el modelo de Hentschel [1995]). La incertidumbre relativa a dicho parámetro, obtenido mediante λ, está cerca de uno y los contrastes de significación fallan en la mayoría de ocasiones. Creemos que una de las posibles fuentes de dicha distorsión está en la capacidad del método propuesto para generar valores extremos. Ello permite concluir que los valores imputados mantienen, en efecto, la leptocurtosis; sin embargo, podría tratarse de un sesgo del método propuesto en la sección *3.3*.

Prospectivas

"Now that you start knowing, pleasure starts flowing"
Mike Pinder, Dawn is a feeling, 1967

I. CON RESPECTO A LA ELECCIÓN Y USO DE **R**

- El desarrollo de la investigación ha confirmado a **R** como una elección adecuada. Sin embargo, la generación de innovaciones mediante el *Gibbs Sampling* y el análisis de los m modelos analizados es una tarea intensiva en tiempo; de ahí nuestra pretensión de implementar el método que hemos llamado de imputación mediante separación lenguajes de ejecución más rápidos como Julia.

II. con respecto a la Imputación múltiple y el número de imputaciones necesarias para su correcta implementación en sección cruzada (inferencia de Rubin)

- La inferencia de Rubin mide la incertidumbre en términos del número de grados de libertad y de ello depende el cálculo de λ, que a su vez está definida de forma individual (una λ para cada escalar de interés). Estas definiciones impiden la medición global de la incertidumbre causada por los NA en el modelo de análisis. Por ello consideramos interesante desarrollar una definición de λ^G o *fracción de información perdida global*.

III. con respecto a la Imputación Múltiple y su correcta implementación en sección longitudinal

- La frecuente infraestimación de los coeficientes relativos a las innovaciones obtenido en el Anexo I y Anexo II, sugiere que el método utilizado para realizar las imputaciones tiene campo para incrementar su eficiencia. Tal vez la mejora de la imputación mediante separación podría hacerse de dos formas: (i) incorporando condiciones previas en el algoritmo MCMC, o (ii) proponiendo factores correctores sobre dichos coeficientes.

- En el capítulo *4*, únicamente se ha estudiado el impacto de los NA dispuestos aleatoriamente. Una investigación pendiente consistiría en el estudio de la plausibilidad de los valores simulados cuando los NA son valores extremos, o cuando pertenecen a un cuantil concreto, o cuando se presentan en fechas determinadas.

- Las pruebas realizadas en los capítulos *3* y *4* se han hecho sobre mercados muy desarrollados y organizados en sentido financiero (sección *4.1*). La investigación debería prolongarse para realizar tests similares en mercados que no cumplieran dichas condiciones; por ejemplo, algunos mercados asiáticos o latinoamericanos. Esto tal vez nos enfrentaría a nuevas problemáticas, que, resueltas, extenderían el procedimiento aquí empleado.

- Una inquietud que tenemos es estudiar cuál sería el resultado del método empleando otros algoritmos. Para simular las innovaciones creemos merece atención el uso del *Data Augmentation* de Tanner & Wong [1987], y para la volatilidad el *bootstrapped EM*.

- En la muestra utilizada, la obtención de colas más pesadas en la distribución condicionada sugiere la necesidad de una investigación más profunda. Este problema tal vez tenga su origen en dos hechos muy distintos: (i) un sesgo del método propuesto o, (ii) aceptando la hipótesis de plausibilidad, la propia naturaleza de las STF genera valores extremos.

- El análisis llevado a cabo en la presente Tesis se ha centrado exclusivamente en modelos GARCH. En el futuro nos agradaría ampliar la gama de modelos analizados usando, por ejemplo, la *Volatilidad Estocástica* de Taylor [1986] o el modelo de carteras propuesto por Markowitz (para el cual podríamos emplear la instrucción `markowitz.portfolio()` contenida en cap. *1*).

Conclusiones globales

A partir de las conclusiones mencionadas hasta aquí opinamos que cabe defender la afirmación de que las imputaciones generadas a partir del método propuesto capturan la esencia de las distintas STF analizadas en esta Tesis, y por lo tanto tal vez sea posible aceptar que los valores generados sean plausibles. Con ello tal vez se hayan alcanzado los objetivos propuestos en la *Introdución*. Sin embargo, la presente investigación abre un campo de análisis que debe ir desarrollándose a través de estudios parciales, quizás siguiendo, en principio, las ideas expuestas en las parte III de las conclusiones prospectivas.

Bibliografía

[1] Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, 67(4):1012–1028.

[2] Alexander, C. (2009). *Market Risk Analysis, Value at Risk Models*, volume 4. Wiley.com.

[3] Alexiadis, M., Dokopoulos, P., Sahsamanoglou, H., and Manousaridis, I. (1998). Short-term forecasting of wind speed and related electrical power. *Solar Energy*, 63(1):61–68.

[4] Allison, P. D. and Oaks, T. (2002). Missing data, quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology*, 55:193–196.

[5] Allison, T. and Cicchetti, D. V. (1976). Sleep in mammals: ecological and constitutional correlates. *Science*, 194(4266):732–734.

[6] Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278):200–203.

[7] Andreu, J. and Cano, S. (2008). Finance forecasting: a multiple imputation approach. *Journal of Applied Mathematics-Aplimat*, 1(1).

[Ardia] Ardia, D. bayesgarch: Bayesian estimation of the garch (1, 1) model with student-t innovations in r, 2007. *URL http://CRAN. R-project. org/package= bayesGARCH.*

[9] Ardia, D. (2008). *Financial risk management with Bayesian estimation of GARCH models: theory and applications*, volume 612. Springer.

[10] Ardia, D. (2009). Bayesian estimation of a markov-switching threshold asymmetric GARCH model with student-t innovations. *Econometrics Journal*, 12(1):105–126.

[11] Ardia, D. (2011). *Bayesian Estimation of the GARCH(1,1) Model with Student-t Innovations in R.* R package version version 1-00.10.

[12] Ardia, D. and Hoogerheide, L. (2010). Bayesian estimation of the garch(1,1) model with student-t innovations. *The R Journal*, 2(2):41–47.

[13] Baillie, R. T. and Bollerslev, T. (1989). The message in daily exchange rates: a conditional-variance tale. *Journal of Business & Economic Statistics*, 20(1):60–68.

[14] Baillie, R. T. and DeGennaro, R. P. (1990). Stock returns and volatility. *Journal of financial and Quantitative Analysis*, 25(2):203–214.

[15] Barnard, J. and Meng, X. L. (1999). Applications of multiple imputation in medical studies: from aids to nhanes. *Statistical Methods in Medical Research*, 8(1):17–36.

[16] Barnard, J. and Rubin, D. (1999a). Miscellanea. small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955.

[17] Barnard, J. and Rubin, D. B. (1999b). Miscellanea. small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955.

[18] Basharin, G., Langville, A., and Naumov, V. (2004). The life and work of a. a. markov. *Linear Algebra and its Applications*, 377.

[19] Beale, E. M. and Little, R. J. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 129–145.

[20] Binder, D. A. and Sun, W. (1996). Frequency valid multiple imputation for surveys with a complex design. In *Proceedings of the Section on Survey Research Methods*, pages 281–286.

[21] Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics*, 3(1):167–179.

[22] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327.

[23] Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics*, 69(3):542–547.

[24] Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). Arch modeling in finance: a review of the theory and empirical evidence. *Journal of econometrics*, 52(1):5–59.

[25] Brooks, C., Burke, S. P., and Persand, G. (2001). Benchmarks and the accuracy of garch model estimation. *International Journal of Forecasting*, 17(1):45–56.

[26] Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 302–306.

[27] Burns, E. M. (1989). Multiple imputation in a complex sample survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 233–238.

[28] Buuren, S. V. and Oudshoorn, C. G. M. (2005). *mice: Multivariate Imputation by Chained Equations*.

[29] Cano, S. and Andreu, J. (2010). Using multiple imputation to simulate time series: A proposal to solve the distance effect. *WSEAS Transactions on Computers*, 9(7):768–777.

[30] Carlin, J. B., Galati, J. C., and Royston, P. (2008). A new framework for managing and analyzing multiply imputed data in stata. *Stata Journal*, 8(1):49–67.

[31] Carlin, J. B., Li, N., Greenwood, P., and Coffey, C. (2003). Tools for analyzing multiple imputed datasets. *The Stata Journal*, 3(3):226–244.

[32] Carpenter, J. (2006). Annotated bibliography on missing data. *accessed July*, 30:2006.

[33] Chen, Q., Ibrahim, J. G., Chen, M. H., and Senchaudhuri, P. (2008). Theory and inference for regression models with missing responses and covariates. *Journal of multivariate analysis*, 99(6):1302–1331.

[34] Chiewchanwattana, S., Lursinsap, C., and Henry Chu, C.-H. (2007). Imputing incomplete time-series data based on varied-window similarity measure of data sequences. *Pattern recognition letters*, 28(9):1091–1103.

[35] Chou, R. Y. (1988). Volatility persistence and stock valuations: Some empirical evidence using garch. *Journal of Applied Econometrics*, 3(4):279–294.

[36] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical statistics*. CRC Press.

[37] Dapugnar, J. S. (2007). *Simulation and Monte Carlo*. Wiley.

[38] Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82(400):1079–1091.

[39] De Jong, P. and Shephard, N. (1995). The simulation smoother for time series models. *Biometrika*, 82(2):339–350.

[40] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

[41] Dempster, A. P. and Rubin, D. B. (1983). Rounding error in regression: The appropriateness of sheppard's corrections. *Journal of the Royal Statistical Society.Series B (Methodological)*, pages 51–59.

[42] Denk, M., Denk, M., and Weber, M. (2011). Avoid filling swiss cheese with whipped cream.

[43] Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(1):49–93.

[44] Ding, Z., Granger, C., and Engle, R. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1):83–106.

[45] Durbin, J. and Koopman, S. J. (2004). *Time series analysis by state space methods*. Oxford University Press.

[46] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26.

[47] Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

[48] Engle, R. and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, 5(1):1–50.

[49] Engle, R., Lilien, D., and Robins, R. (1987). Estimating time varying risk premia in the term structure: The arch-m model. *Econometrica: Journal of the Econometric Society*, 55(2):391–407.

[50] Engle, R. and Ng, V. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance*, 48(5):1749–1778.

[51] Engle, R. F. and Sokalska, M. E. (2012). Forecasting intraday volatility in the us equity market. multiplicative component garch. *Journal of Financial Econometrics*, 10(1):54–83.

[52] Ferreiro, O. (1987). Methodologies for the estimation of missing observations in time series. *Statistics & probability letters*, 5(1):65–69.

[53] Gallant, A. R. and Tauchen, G. (1989). Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications. *Econometrica: Journal of the Econometric Society*, pages 1091–1120.

[54] Gamerman, D. and Lopes, G. (2006). *Markov Chain Monte Carlo*. Chapman and Hall.

[55] García, J. C. F., Kalenatic, D., and Bello, C. A. L. (2008). Missing data imputation in time series by evolutionary algorithms. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, pages 275–283. Springer.

[56] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). Bayesian data analysis. 2004.

[57] Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

[58] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian distribution of images. *IEEE Trans. Pattern Anal. Machine Intell*, 6(6):721–741.

[59] Geweke, J. (1993). Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics*, 8(S1):S19–S40.

[60] Ghalanos, A. (2013). *rugarch: Univariate GARCH models.* R package version 1.0-16.

[61] Glosten, L., Jagannathan, R., and Runkle, D. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801.

[62] Goodman, J. and Sokal, A. D. (1989). Multigrid monte carlo method. conceptual foundations. *Physical Review D*, 40(6):2035–2071.

[63] Goulet, V., Dutang, C., Maechler, M., Firth, D., Shapira, M., Stadelmann, M., and expm-developers@lists.R-forge.R-project.org (2012). *expm: Matrix exponential*. R package version 0.99-0.

[64] Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576.

[65] Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213.

[66] Grinstead, C. and Snell, L. (2006). *Grinstead and Snell's Introduction to probability*. American Mathematical Society.

[67] Groves, R. M. (1989). *Survey errors and survey costs*, volume 536. Wiley.

[68] Gujarati, D. (1997). Econometría, ed. *McGraw Hill*.

[69] Hartley, H. and Hocking, R. (1971). The analysis of incomplete data. *Biometrics*, 27(4):783–823.

[70] Harvey, A. C. and Pierse, R. G. (1984). Estimating missing observations in economic time series. *Journal of the American Statistical Association*, 79(385):125–131.

[71] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97.

[72] He, Y., Yucel, R., and Raghunathan, T. E. (2011). A functional multiple imputation approach to incomplete longitudinal data. *Statistics in Medicine*, 30(10):1137–1156.

[73] Healy, M. and Westmacott, M. (1956). Missing values in experiments analysed on automatic computers. *Applied statistics*, pages 203–206.

[74] Heeringa, S. (1993). Imputation of item missing data in the health and retirement survey. *Invited paper.Proceedings of the Survey Methods Section*, pages 107–116.

[75] Heitjan, D. F. and Little, R. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics*, 40(1):13–29.

[76] Heitjan, D. F. and Rubin, D. B. (1990). Ignorability and coarse data. *The Annals of Statistics*, pages 2244–2253.

[77] Hentschel, L. (1995). All in the family nesting symmetric and asymmetric garch models. *Journal of Financial Economics*, 39(1):71–104.

[78] Herzog, T. and Lancaster, C. (1980). Multiple imputation of individual social security amounts.

[79] Herzog, T. N. (1980). Multiple imputation modeling for individual social security benefit amounts, part ii. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 404–407.

[80] Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2):561–581.

[81] Honaker, J., King, G., Blackwell, M., et al. (2006). Amelia ii: A program for missing data.

[82] Horton, N. and Lipsitz, R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statiscian*, 55(3):244–254.

[83] Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing. *The American Statistician*, 61(1):79–90.

[84] Horton, N. J., Lipsitz, S. R., and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, 57(4):229–232.

[85] Hsieh, D. A. (1989). Modeling heteroscedasticity in daily foreign-exchange rates. *Journal of Business & Economic Statistics*, 7(3):307–317.

[86] Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769.

[87] Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models. *Journal of the American Statistical Association*, 100(469):332–346.

[88] Jackman, S. (2000). Estimation and inference via bayesian simulation: An introduction to markov chain monte carlo. *American Journal of Political Science*, 44(2):375–404.

[89] Jarrett, R. G. (1978). The analysis of designed experiments with missing observations. *Applied Statistics*, pages 38–46.

[90] Jung, H., Schafer, J. L., and Seo, B. (2010). A latent class selection model for nonignorably missing data. *Computational Statistics & Data Analysis*.

[91] Kalman, R. E. et al. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

[92] Kennedy, A. and Kulti, J. (1985). Noise without noise: A new monte carlo method. *Physical Review Letters*, 54(23):2473–2476.

[93] Kennickell, A. B. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 1–10.

[94] Kennickell, A. B. (1999). *Multiple Imputation and Disclosure Control: The Case of the 1995 Survey of Consumer Finances*. Record Linkage Techniques 1997.

[95] Kenward, M. G. and Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16(3):199.

[96] Khare, M., Little, R. J. A., Rubin, D. B., and Schafer, J. L. (1993). Multiple imputation of nhanes 3 (with discussion). *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 297–302.

[97] Kihoro, J., Otieno, R., and Wafula, C. (2007). Seasonal time series data imputation: Comparison between feed forward neural networks and parametric approaches. *East African Journal of Statistics*, 1(1):68–83.

[98] Lange, K. (1999). *Numerical Analysis for Statisticians*. Springer.

[99] Lee, G. and Engle, R. (1999). A permanent and transitory component model of stock return volatility. In *Cointegration Causality and Forecasting A Festschrift in Honor of Clive WJ Granger*, pages 475–497. Oxford University Press.

[100] Li, K. H. (1988). Imputation using markov chains. *Journal of Statistical Computation and Simulation*, 30(1):57–79.

[101] Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, pages 287–296.

[102] Little, R. J. A. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, pages 1227–1237.

[103] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons.

[104] Mandelbrot, B. (1967). The variation of some other speculative prices. *Journal of Business*, 40(4):393–413.

[105] Mandelbrot, B. B. (1963). *The variation of certain speculative prices*. Springer.

[106] Markov, A. (2006). An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Science in Context*, 19:591–600.

[107] Markowitz, H. (1952). Portfolio selection. *The journal of finance*, 7(1):77–91.

[108] McCullough, B. and Renfro, C. G. (1999). Benchmarks and software standards: A case study of garch procedures. *Journal of Economic and Social Measurement*, 25(2):59–71.

[109] McKendrick, A. G. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130.

[110] Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558.

[111] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087.

[112] Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247):335–341.

[113] Nakatsuma, T. (1998). A markov-chain sampling algorithm for garch models. *Studies in Nonlinear Dynamics and Econometrics*, 3(2):107–117.

[114] Neal, R. M. (1993). Probabilistic inference using markov chain monte carlo methods. Technical report, University of Toronto.

[115] Nelson, D. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–70.

[116] Oh, H. and Scheuren, F. (1980). Estimating the variance impact of missing cps income data. *Proceedings of the Section on Survey Research Methods*, pages 408–415.

[117] Pagan, A. R. and Schwert, G. W. (1990). Alternative models for conditional stock volatility. *Journal of Econometrics*, 45(1):267–290.

[118] Pan, W. (2001). A multiple imputation approach to regression analysis for doubly censored data with application to aids studies. *Biometrics*, 57(4):1245–1250.

[119] Parzen, E. (1984). Time series analysis of irregularly observed data. In *Lecture Notes in Statistics, Proceedings of a Symposium, held at College Station, Texas, USA, February 10-13, 1983, New York: Springer, 1984, edited by Parzen, Emanuel*, volume 1.

[120] Pearce, S. C. et al. (1965). *Biological statistics: an introduction*. Cambridge Univ Press.

[121] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

[122] Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

[123] Pfaff, B. (2012). *gogarch: Generalized Orthogonal GARCH (GO-GARCH) models*. R package version 0.7-2.

[124] Poon, S.-H. and Taylor, S. J. (1992). Stock returns and volatility: an empirical study of the uk stock market. *Journal of banking & finance*, 16(1):37–59.

[125] Preece, D. (1971). Iterative procedures for missing values in experiments. *Technometrics*, 13(4):743–753.

[126] R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[127] Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96.

[128] Rassler, S. (2002). *Statistical Matching*. Springler.

[129] Redman, T. C. (1992). *Data quality: management and technology*. Bantam Books, Inc.

[130] Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika*, 94(2):502.

[131] Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471.

[132] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[133] Royston, P. (2005). Multiple imputation of missing values: update. *Stata Journal*, 5(2):188.

[134] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581.

[135] Rubin, D. B. (1978). Multiple imputations in sample surveys: a phenomenological bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, pages 20–34.

[136] Rubin, D. B. (1981). The bayesian bootstrap. *The annals of statistics*, 9(1):130–134.

[137] Rubin, D. B. (1987). *Multiple imputation for nonresponse in survey*. Wiley.

[138] Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91.

[139] Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394):366–374.

[140] Rubin, G. (2006). The traffic in women: Notes on the political economy of sex. *Feminist Anthropology: a reader*.

[141] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall / CRC Press.

[142] Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:3–15.

[143] Schafer, J. L., Ezzatti-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., and Rubin, D. B. (1996). The nhanes 3 multiple imputation project. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 28–27.

[144] Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(147-177).

[145] Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems. *Multivariate Behavioral Research*, pages 545–571.

[146] Schenker, N. and Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation* 1. *Computational statistics & data analysis*, 22(4):425–446.

[147] Schenker, N., Treiman, D. J., and Weidman, L. (1993). Analyses of public use decennial census data with multiply imputed industry and occupation codes. *Applied Statistics*, 42:545–556.

[148] Schwert, G. (1990). Stock volatility and the crash of '87. *Review of Financial Studies*, 3(1):77.

[149] Siegl, T. and Quell, P. (2005). Modelling specific interest rate risk with estimation of missing data. *Applied Mathematical Finance*, 12(3):283–309.

[150] Steele, R. J., Wang, N., and Raftery, A. E. (2010). Inference from multiple imputation for missing data using mixtures of normals. *Statistical methodology*, 7(3):351–365.

[151] Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, pages 49–58.

[152] Swendsen, R. H. and Wang, J. S. (1987). Nonuniversal critical dynamics in monte carlo simulations. *Physical Review Letters*, 58(2):86–88.

[153] Szeliski, R. (1989). *Bayesian Modeling of Uncertaintyin Low-Level Vision*. Boston: Kluwer.

[154] Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

[155] Taylor, S. (1986). *Modelling financial time series*. Wiley.

[156] Team, R. (2010). R: A language and environment for statistical computing. *R Foundation for Statistical Computing Vienna Austria*, (01/19).

[157] Trapletti, A. and Hornik, K. (2012). *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-30.

[158] Vach, W. (1994). *Logistic regression with missing values in the covariates*. Springer-Verlag New York.

[159] Vach, W. and Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology*, 134(8):895.

[160] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

[161] Wuertz, D., Chalabi, Y., Chen, W., and Ellis, A. (2010). *Portfolio Optimization with R/Rmetrics*. Rmetrics Association Finance Online, www.rmetrics.org. R package version 2130.80.

[162] Wuertz, D., with contribution from Michal Miklovic, Y. C., Boudt, C., Chausse, P., and others (2012). *fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*. R package version 2150.81.

[163] Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture*, 1(2):129–142.

[164] Zakoian, J. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control*, 18(5):931–955.