Tesis de Master

Funciones de autocorrelación robustas, test estacionarios y test de raíces unitarias

Juan Luis Vicente Calvo

Funciones de autocorrelación robustas, test estacionarios y test de raíces unitarias

Universidade Aberta de Portugal
Universidad Nacional de Educación a Distancia

Juan Luis Vicente Calvo

Tesis de Master: **30 ECTS**

Tutora: **Maria do Rosário Ramos**

Septiembre de 2010

Índice de tablas

Índice General

Introducción

Un requerimiento básico para el modelado de series temporales es que las series en estudio deben ser débilmente estacionarias. Numerosos test han sido desarrollados y popularmente aplicados, en especial el test de Dickey-Fuller de raíces unitarias.

En un test de raíces unitarias, la hipótesis nula es aquella en la que el proceso contiene una raíz unitaria mientras que la alternativa es aquella en la que es estacionario. Por otro lado los test que buscan la estacionariedad operan en la dirección opuesta. La hipótesis nula es que la serie es estacionaria mientras que la alternativa es que existe una componente no estacionaria. La distribución asintótica del estadístico bajo la hipótesis nula de estacionariedad es la distribución de Cramér-von Mises.

Kwiatkowski, Phillips, Schmidt y Shim proponen un test, de raíces unitarias, basado en la métrica de Cramér-von Mises, el test estadístico llamado KPSS (1992). Además, se puede utilizar la métrica de Cramér-von Misses o la de Kolmogorov-Smirnov para contrastar la hipótesis nula que especifica, por ejemplo, que la función de distribución espectral, que es una función lineal de las autocorrelaciones, sigue una completamente especificada y estacionaria, el estadístico es una integral de una función cuadrática de un proceso estocástico gaussiano y la distribución asintótica del test puede ser aproximada por una suma infinita, Anderson, T.W . (1993,1995) , Anderson, T.W y Stephens, M.A. (1993) muestran como calcular puntos asintóticos del test.

Por otro lado sabemos que, convencionalmente, el test de Dickey-Fuller es sesgado en presencia de outliers aditivos (OA). Los valores críticos son muy sensibles a el tipo de ruptura, el tiempo y la magnitud de la ruptura. Sin embargo se pueden utilizar correcciones de los valores críticos introduciendo variables Dummy en la obtención de la regresión y por

tanto, desde un punto de vista empírico, se pueden obtener resultados utilizando una conveniente selección de variables dummy.

De igual forma, el test KPSS depende funcionalmente de la función de autocorrelaciones así como el test de bondad de ajuste de Anderson, en este la función de distribución espectral muestral y teórica son funciones lineales de las autocorrelaciones y por tanto no robustas en presencia de outliers aditivos.

En este trabajo estudiamos el problema de raíces unitarias en presencia de outliers aditivos desde un punto de vista robusto y se sugiere el uso de un procedimiento robusto de estimación de la función de autocorrelaciones.

Bajo este contexto se pueden utilizar estimadores robustos con alto punto de ruptura, el parámetro de localización multivariado permanece acotado y los valores propios de la matriz de varianza-covarianza permanecen lejos de cero y de infinito.

Sin embargo la definición de punto de ruptura pierde su significado si trabajamos con series temporales y funciones de autocorrelación, por un lado la perturbación no es igual si se encuentra al inicio o en medio, además el efecto dependerá notablemente de la distancia del retardo. En esta situación es necesario introducir una nueva definición el punto de ruptura temporal de un estimador de autocovarianza y esta definición es dada por Genton y Ma (1998) .

Utilizando funciones de autocorrelación robustas proponemos un estimador basado en el algoritmo Bagging para clasificación, Breiman (1999). Esta metodología utiliza tanto técnicas de perturbación como de combinación para reducir la variabilidad en estimadores inestables. Al igual que en Random Forest, Breiman (1999), introducimos una variante en el algoritmo Bagging.

En series temporales existen dos vías para submuestrear en secuencias de datos dependientes, basado en modelos y libres de modelo. En el tipo basado en modelos la estructura de dependencia es modelada en términos de un conjunto de parámetros desconocidos y errores independientes. En el tipo libre de modelo la serie observada es dividida en bloques y, estos bloques son usados para capturar la dependencia en la serie original.

Finalmente se presenta por simulación de 2000 muestras, basadas en diferentes tamaños, una comparación de los diferentes test, el test de Dickey-Fuller clásico, el test basado en métodos de estimación robustos y el test basado en aprendizaje estadístico

utilizando métodos de estimación robustos. Como criterio de comparación se utiliza la función potencia del test. Para todas las simulaciones se han realizado una serie de procedimientos en el software estadístico SAS.

1. Raíces unitarias y AO

1.1 Introducción

El modelo de regresión que se va a considerar es el caso con errores autorregresivos de primer orden o $AR(1)$, en esta situación la serie univariada $\{y_t; t = 1,\ldots,T\}$, es generada por

$$y_t = \mu + \beta t + u_t;\ u_t = \phi u_{t-1} + \varepsilon_t,$$

donde $\varepsilon_t \approx i.i.d.(0,\sigma^2)$ y $E(\varepsilon_t^4) < \infty$. Aquí supondremos que $-1 \le \phi \le 1$ y en tal caso procesos estacionarios e integrados están permitidos.

Así pues si consideramos un proceso $AR(1)$, $U = (U_t : t \in \aleph := \{0,1,2,\ldots\})$, el cual satisface la siguiente ecuación en diferencias estocástica

$$U_t = \phi U_{t-1} + \varepsilon_t\ ,\quad t \in \aleph\ ,\quad \phi \in \Re\ ,$$

donde el proceso de ruido blanco, $(\varepsilon_t : t \in \aleph = \{1,2,\ldots\})$, es una secuencia de variables aleatorias independientes e idénticamente distribuidas con media nula y varianza, $\sigma^2 > 0$, y función de distribución F.

En el caso de que $|\phi| < 1$ entonces existe para un proceso ruido blanco una solución estrictamente estacionaria $(U_t : t \in \aleph)$ y en tal caso se cumple que

$$\zeta(U_{t_1},\ldots,U_{t_r}) = \zeta(U_{t_1+h},\ldots,U_{t_r+h})$$

para todo $t_1,\ldots,t_r, r, h \in \aleph$. Donde $\zeta(.)$ denota la distribución de un vector aleatorio. Asumiremos que si $U_0 = 0$ y si $|\phi| < 1$ entonces estaremos ante el caso estacionario. Por

otro lado si $|\phi|>1$ el proceso será llamado explosivo y no es estacionario. Y finalmente para el caso $\phi=\pm 1$ que también será un proceso no estacionario y será llamado inestable.

El estimador de mínimos cuadrados que depende de las primeras n observaciones tanto para el caso estacionario como no estacionario es

$$\hat{\phi}_n=\frac{\sum_{t=2}^{n}U_tU_{t-1}}{\sum_{t=2}^{n}U^2{}_{t-1}},$$

y por tanto tenemos que

$$\begin{array}{lll} si & |\phi|<1, & \sqrt{n}\left(\hat{\phi}_n-\phi\right) \rightarrow \mathrm{N}\left(0,1-\phi^2\right), \quad n\rightarrow\infty \\ si & |\phi|=1, & n\left(\hat{\phi}_n-\phi\right) \rightarrow L(\phi), \qquad\quad n\rightarrow\infty, \\ si & |\phi|>1, & |\phi|^n\left(\hat{\phi}_n-\phi\right) \rightarrow Q(\phi), \qquad\quad n\rightarrow\infty \end{array}$$

donde la distribución límite $L(\phi)$ para el caso $|\phi|=1$ fue derivada por Rao (1978). Para el caso $\phi=1$ tenemos que

$$L(1)=\frac{1}{2}\left(W^2(1)-1\right)\left(\int_0^1 W^2(s)ds\right)^{-1},$$

donde W es un proceso estándar de Wiener. Y para el caso $\phi=-1$

$$L(-1)=\frac{1}{2}\left(1-W^2(1)\right)\left(\int_0^1 W^2(s)ds\right)^{-1}.$$

Anderson (1959) ha mostrado que existen dos variables aleatorias independientes e idénticamente distribuidas H y V, tal que

$$Q(\phi)=\left(\phi^2-1\right)\frac{V}{H}.$$

Si ε_1 esta distribuida como una variable aleatoria normal entonces H y V están distribuidas normalmente con media nula y varianza $\sigma^2\phi^2/\left(\phi^2-1\right)$,

$$L(H)=L(V)=L\left(\sum_{j=1}^{\infty}\phi^{1-j}\varepsilon_j\right).$$

Por otro lado, de igual forma si $\varepsilon_t \approx NID(0,1)$ y consideramos la función de densidad conjunta

$$f_\phi(u)=\left(\sqrt{2\pi}\right)^{-n}\exp\left\{-\frac{1}{2}\left[(1+\phi)^2T_2-2\phi T_1+u_n^2\right]\right\},$$

donde $T_1 = \sum_{t=1}^{n} u_t u_{t-1}$, $T_2 = \sum_{t=1}^{n} u_{t-1}^2$. El estimador de máximaverosimilitud de ϕ es dado por

$$\hat{\phi}_n = \frac{T_1}{T_2}.$$

Entonces dado $u_0 = 0$

$$\sqrt{I(\phi)}\left(\hat{\phi} - \phi\right) \xrightarrow{d} \begin{cases} N, & |\phi| < 1 \\ \phi \dfrac{W^2(1) - 1}{2^{3/2} \int_0^1 W^2(r) dr}, & |\phi| = 1 \\ C, & |\phi| > 1 \end{cases}$$

donde N es una variable aleatoria Gausiana centrada con varianza 1, C es una variable aleatoria de Cauchy, $W = \left(W(r), r \geq 0\right)$ es un proceso de Wiener y $I_n(\phi)$ es la información de Fisher contenida en la muestra sobre el parámetro ϕ dado $E(T_2)$. Cuando $n \to \infty$, Mann and Wald White, Anderson, Dickey y Fuller muestran que

$$E(T_2) \approx \begin{cases} \dfrac{n}{1 - \phi^2}, & |\phi| < 1, \\ \dfrac{n^2}{2}, & |\phi| = 1 \\ \dfrac{\phi^{2n}}{\left(\phi^2 - 1\right)^2}, & |\phi| > 1 \end{cases},$$

por tanto si $|\phi| \leq 1$ el resultado es valido asumiendo que u_0 es una constante arbitraria o una variable aleatoria con momentos de segundo orden finitos e independientes de $\varepsilon_t, t \geq 1$, siendo, en esta situación, una sucesión de variables aleatorias centradas independientes e idénticamente distribuidas.

Si $|\phi| > 1$, la distribución límite depende del valor inicial y, en general de la distribución de ε_t, incluso si forman una sucesión de variables aleatorias independientes e idénticamente distribuidas.

Bajo este contexto si normalizamos por el valor esperado de la información de Fisher dado $x_0 = 0$, entonces

$$\sqrt{T_2}\left(\hat{\phi} - \phi\right) \xrightarrow{d} \begin{cases} N, & |\phi| \neq 1 \\ \phi \dfrac{W^2(1) - 1}{2^{3/2} \int_0^1 W^2(r) dr}, & |\phi| = 1 \end{cases}$$

y sería posible obtener una única distribución límite por estimación secuencial.

Por su parte Mikulski and Monsour (1991) consideran la clase de estimadores tales que dado ϕ, el sesgo es diferenciable en ϕ y satisface las condiciones

$$b_\phi(\phi_n) \to 0, \frac{db_\phi(\phi_n)}{d\phi} \to 0, \quad cuando \quad n \to \infty,$$

entonces el estimador pertenece a la clase de estimadores para la cual $|\phi| \neq 1$ y si $|\phi| \neq 1$ entonces el estimador de máximaverosimilitud es asintóticamente eficiente en esta clase. Si $|\phi| < 1$, el estimador es eficiente. Finalmente si $|\phi| = 1$ el estimador no pertenece a esta clase de estimadores y, claro esta, estamos ante el caso inestable.

Por tanto bajo este contexto no existen expresiones exactas o aproximadas para calcular funciones de densidad y el camino que se utiliza es derivar funcionales de Wiener y simular valores críticos. Así pues para un camino aleatorio

$$U_t = U_{t-1} + \varepsilon_t, \quad t \in \aleph$$

la serie es no estacionaria, claramente $V(u_t) = V(u_{t-1}) + \sigma^2$, la distribución límite es un movimiento Browniano. Dado $u_0 = 0$ y definiendo el proceso continuo $u_w^{(n)}(t): [0,1] \to \Re$ conectando los puntos $\left(t/n, u_t/\sqrt{n}\right)$. Entonces si $E\left(|\varepsilon|_t^3\right) < \infty$, utilizando el teorema central del límite, Borovkov (1999), cuando $n \to \infty$ se tiene que

$$\frac{1}{\sigma} u_w^{(n)}(t) \to W(t),$$

y en esta situación el resultado puede usarse para encontrar distribuciones asintóticas de test de raíces unitarias.

Se puede utilizar, por ejemplo, la distribución de Cramer-von Misses para contrastar la hipótesis nula que especifica que la función de distribución espectral sigue una completamente especificada. Para un camino aleatorio la función de densidad espectral muestral es dada por

$$I(\lambda) = \frac{\sigma^2}{2\pi} \frac{1}{\sin^2 \frac{\lambda}{2}} \left[(n+1) - \frac{\sin \frac{(n+1)\lambda}{2}}{\sin \frac{\lambda}{2}} \right].$$

De hecho si consideramos el modelo autorregresivo de primer orden y definimos la hipótesis nula de estacionariedad contra la alternativa de no estacionariedad, bajo esta situación el test bajo la hipótesis nula puede formularse como

$$\lambda = \frac{T^{-2}\sum_{i=1}^{T}\left[\sum_{t=1}^{i}\varepsilon_t\right]^2}{s^2} > c,$$

donde $s^2 = T^{-1}\sum_{t=1}^{T}(u_t - \bar{u})^2$ y c es el llamado valor crítico del test.

La distribución asintótica del test bajo la hipótesis nula puede encontrarse observando que las desviaciones en sumas parciales sobre la media convergen a un puente browniano o brownian bridge, luego

$$\sigma^{-1}T^{-1/2}\sum_{s=1}^{[Tr]}\varepsilon_s \rightarrow B(r), \quad r \in [0,1],$$

donde $[Tr]$ es el más grande entero menor o igual a Tr y $B(r) = W(r) - rW(1)$. Entonces

$$\lambda \rightarrow \int_0^1 B(r)^2\,dr\,,$$

que no es sino la distribución de Cramér-von Mises.

1.2 Estimador de mínimos cuadrados y AO

Supongamos el modelo con outliers aditivos (AO) introducido por Fox (1972), el modelo consiste de un proceso estacionario u_t el cual es observado con error $\tilde{u}_t = u_t + z_t$ siendo

$$z_t = \begin{cases} \varepsilon & si\ t = t_0 \\ 0 & otro\ caso \end{cases},$$

y la contaminación, z_t, son variables aleatorias independientes e idénticamente distribuidas con distribución dada por $F_z = (1-\varepsilon)\delta_0 + \varepsilon H$ donde, δ_0 es una distribución degenerada teniendo toda su masa en el origen y H es una distribución con media cero y varianza σ_H^2.

El estimador de mínimos cuadrados de ϕ basado en la muestra contaminada $\tilde{u}_1,\ldots,\tilde{u}_T$ es

$$\hat{\phi} = \frac{\sum_{t=1}^{T} \tilde{u}_t \tilde{u}_{t-1}}{\sum_{t=2}^{T} \tilde{u}_{t-1}} = \frac{\varepsilon\left(u_{t_0-1} + u_{t_0+1}\right) + \sum_{t=2}^{T} \tilde{u}_t \tilde{u}_{t-1}}{\varepsilon^2 + 2\varepsilon u_{t_0} + \sum_{t=2}^{T} \tilde{u}_{t-1}^2}.$$

En esta situación se observa que cuando $\varepsilon \to \infty$ entonces $\hat{\phi} \to 0$ y la estimación esta totalmente determinada por la contaminación. En consecuencia podemos decir que el estimador de ϕ rompe a 0 con un solo dato anómalo. Lo más característico es que el estimador no diverge a ∞ y por tanto, esta forma de ruptura, no esta acorde con la definición clásica de Hampel (1968) y la dada por Hodges (1967), que definió el punto de ruptura, para tamaños muestrales finitos, como la proporción más pequeña de las n observaciones que puede hacer no acotado superiormente el estimador. Otras extensiones a tamaños muestrales finitos son las dadas por Donoho, y Huber (1983) y, para observaciones dependientes, Martin y De Jong, (1977); Martin (1980); Genton (1988); Ma y Genton, (2000). Existen otras definiciones de punto de ruptura, por ejemplo la de He y Simpson (1992), (1993), en estas se busca el punto de ruptura si el supremo del sesgo es alcanzado. Sin embargo, aquí el supremo del sesgo se alcanza cuando $\hat{\phi}$ tiende a 1 o -1, dependiendo del signo del parámetro y por tanto tampoco estaría acorde a esta definición. En esta situación es necesario definir una medida de ruptura, y es aquella que ha sido dada por Genton y Ma (2000).

1.3 Test de Dickey-Fuller y AO

En un test de raíces unitarias, la hipótesis nula es aquella en la que el proceso contiene una raíz unitaria mientras que la alternativa es aquella en la que es estacionario. La distribución asintótica de un test de raíces unitarias $T\left(\hat{\phi}_T^{MCO} - 1\right)$ viene dada por una composición de Movimientos Brownianos, en efecto si

$$T\left(\hat{\phi}_T^{MCO} - 1\right) = \frac{T^{-1}\sum_{t=1}^{T} u_{t-1} a_t}{T^{-2}\sum_{t=1}^{T} u_{t-1}^2},$$

entonces

$$T^{-1}\sum_{t=1}^{T} u_{t-1}a_t \rightarrow \frac{1}{2}\sigma^2\left[W(1)^2 - 1\right],$$

y

$$T^{-2}\sum_{t=1}^{T} u_{t-1}^2 \rightarrow \sigma_u^2 \int_0^1 W(1)^2\, dr\,,$$

por tanto

$$T\left(\hat{\phi}_T^{MCO} - 1\right) = \frac{1/2\left[W(1)^2 - 1\right]}{\int_0^1 W(r)^2\, dr}.$$

Ahora si consideramos que u_t es un proceso generado por un camino aleatorio con inicio en el origen, $u_0 = 0$

$$u_t = u_{t-1} + \varepsilon_t \quad t = 1,2,\ldots,T,$$

donde ε_t son variables aleatorias independientes e idénticamente distribuidas , $N\left(0,\sigma_\varepsilon^2\right)$, si suponemos outliers aditivos de tamaño $\pm\theta$, los cuales ocurren con probabilidad π, entonces la serie observada es

$$x_t = u_t + \theta\sigma_t\,,$$

donde σ_t son variables aleatorias Bernoulli, de forma que

$$P(\sigma_t = 1) = 1/2\pi,\ P(\sigma_t = -1) = 1/2\pi\,,$$

y

$$P(\sigma_t = 0) = 1 - \pi,\ \sigma_0 = 0\,.$$

Si

$$\Delta u_t = \phi u_{t-1} + \varepsilon_t\,,$$

el test de Dickey-Fuller de raíces unitarias toma como hipótesis nula $H_0 : \phi = 0$, de raíz unitaria, frente a la alternativa $H_1 : \phi < 0$, de raíz estable. El test es implementado por medio del estimador de mínimos cuadrados de ϕ, el cual es

$$\hat{\phi} = \frac{\sum_{t=1}^{T} u_{t-1}\Delta u_t}{\sum_{t=1}^{T} u_{t-1}^2}\,,$$

y el correspondiente t estadístico para un coeficiente cero es

$$t_{\hat{\phi}} = \frac{\hat{\phi}}{\hat{\sigma}_u \left(\sum_{t=1}^{T} u_{t-1}^2 \right)^{1/2}},$$

donde $\hat{\sigma}_u^2 = T^{-1} \sum_{t=1}^{T} \left(\Delta u_t - \hat{\phi} u_{t-1} \right)^2$, $\Delta = (1 - L)$ y L es el operador retardo.

En presencia de AO tenemos que

$$\Delta x_t = \phi x_{t-1} + v_t$$

donde $v_t = \varepsilon_t + \theta \Delta \delta_t$ y el estimador de mínimos cuadrados correspondiente es

$$\hat{\phi}_{AO} = \frac{\sum_{t=1}^{T} x_{t-1} \Delta x_t}{\sum_{t=1}^{T} x_{t-1}^2}$$

y

$$t_{\hat{\phi}_{AO}} = \frac{\hat{\phi}_{AO}}{\hat{\sigma}_x \left(\sum_{t=1}^{T} x_{t-1}^2 \right)^{-1/2}},$$

donde $\hat{\sigma}_x^2 = T^{-1} \sum_{t=1}^{T} \left(\Delta x_t - \hat{\phi}_{AO} x_{t-1} \right)^2$. Se deduce que bajo la hipótesis nula, H_0, la serie observada es un proceso $I(1)$ con errores $MA(1)$ o proceso de medias móviles de orden 1, dado que $E[v_t v_{t-1}] = -\pi\theta^2$, $E[v_t v_{t-j}] = 0$ *para* $j > 1$. Bajo esta situación v_t satisface las condiciones de Phillips (1987) y, si utilizamos el teorema funcional central del límite aplicado a las sumas parciales de v_t para obtener la distribución asintótica, cuando $T \to \infty$

$$T_{\hat{\phi}AO} \to \left(\int_0^1 W(r) dW(r) \right) \Big/ \left(\int_0^1 W^2(r) dr \right) - (\theta/\sigma_\varepsilon)^2 \pi \left(\int_0^1 W^2(r) dr \right)^{-1}$$

y

$$t_{\hat{\phi}AO} \to \left(1 + 2(\theta/\sigma_\varepsilon)^2 \pi\right)^{-1/2} \left(\int_0^1 W(r) dW(r) \right) \Big/ \left(\int_0^1 W^2(r) dr \right)^{1/2},$$

donde $W(r)$ es un movimiento browniano estándar definido en el intervalo $r \in [0,1]$.

Estas ecuaciones fueron derivadas por Franses y Haldrup (1994). Se puede deducir que si $\pi > 0$, el parámetro ϕ es todavía estimado superconsistentemente pero la distribución asintótica del estadístico de Dickey-Fuller será sesgada a la derecha, consecuentemente se sobre rechazara la hipótesis de raíz unitaria a favor de la alternativa

estacionaria. Notar que un AO positivo y un AO negativo tienen la misma magnitud en el efecto en la distribución límite. También cabe destacar que las distribuciones son iguales para cualquier combinación de θ, π y σ_ε^2 dando el mismo valor de $(\theta/\sigma_\varepsilon)^2\pi$. En otras palabras, largos shocks con pequeña probabilidad de ocurrencia tienen el mismo efecto que pequeños shocks con alta probabilidad de ocurrencia. Por tanto el test de Dickey-Fuller no puede distinguir infrecuentes largos shocks de frecuentes pequeños.

2. Funciones de autocorrelación

2.1 Introducción

Test de bondad de ajuste así como los test de raíces unitarias para series temporales se basan en la secuencia de autocorrelaciones o bien dependen funcionalmente de la función de autocovarianzas o de la función autocorrelaciones.

Por otro lado nos encontramos test basados en la representación espectral del proceso estocástico estacionario, en este contexto la representación de $\{y_t\}$ puede hacerse en términos de integrales de procesos estocásticos, estas integrales representan la función de distribución espectral, función de varianzas de amplitudes aleatorias de funciones trigonométricas que componen el proceso. Anderson, T.W . (1993,1995) , Anderson, T.W y Stephens, M.A. (1993) utilizan el hecho de que la densidad espectral muestral puede ser utilizada para dar una estimación de la densidad espectral.

La función de distribución espectral muestral $F_T(v) = \int_{-\pi}^{\lambda} I(\lambda) d\lambda$ es un estimador de la función de distribución espectral y la diferencia $\sqrt{T}\left[\hat{F}_T(\lambda) - F(\lambda)\right]$ tiene distribución asintóticamente normal, esta diferencia es un proceso estocástico en $[0, \pi]$. Test como el Cramer-von Misses o test de Kolmogorov-Smirnov están basados en esta diferencia. En esta situación la función de covarianza asintótica de esta diferencia puede ser encontrada.

Por tanto se puede utilizar el test de Cramer-von Misses para contrastar la hipótesis nula que especifica que la función de distribución espectral sigue una completamente especificada, el estadístico es una integral de una función cuadrática de un proceso

estocástico gaussiano, la distribución asintótica del test puede ser aproximada por una suma infinita

$$W^2 = \sum_{i=1}^{\infty} w_i X_i^2$$

donde $\{X_i\}$ son variables aleatorias idénticamente distribuidas $N(0,1)$, Anderson y Stephens (1993,1995), muestran como calcular los pesos w_i utilizando métodos numéricos y a partir de ellos dan puntos asintóticos del test, estos son dados para un $AR(1)$. Este procedimiento resulta demasiado complejo para aplicarlo a otros procesos aunque es descrito para un proceso autorregresivo de orden p, en este caso $f(\lambda)$ o equivalentemente $F(\lambda)$ dependen de un vector de parámetros, estos son los coeficientes asociados al proceso lineal. La distribución límite del estadístico es obtenida.

El criterio de Kolmogorov-Smirnov, aunque más conocido que el de Cramer-von Mises, no es tan potente como este, una aproximación a este para series temporales es encontrada a partir del hecho de que $\sqrt{T}\left[\hat{F}_T(\lambda) - F(\lambda)\right]$ sigue la distribución asintótica de un puente browniano o brownian bridge.

2.2 Función de autocorrelación

Dada una muestra aleatoria $y_1, \ldots, y_T$ estamos interesados en contrastar la hipótesis nula que especifica que la función de autocorrelaciones del proceso sigue un determinado patrón. Un proceso estocástico estacionario, si es gaussiano, esta completamente determinado por la media y su función de autocorrelaciones. La hipótesis alternativa vendrá dada por todos los patrones no especificados en la hipótesis nula.

En el dominio de la frecuencia, la función de distribución espectral muestral y la teórica son funciones lineales de las autocorrelaciones, luego podemos utilizar estas para construir test de bondad de ajuste para comparar si la función de distribución espectral de un proceso estocástico sigue una totalmente especificada. La función de distribución espectral muestral puede utilizarse como estimación de la función de distribución espectral. Entre ellos el test de Cramér-von Mises o test de Kolmogorov-Smirnov.

Un proceso estocástico $\{y_t\}$, $t = \ldots, -1, 0, 1, \ldots$, en tiempo discreto, estacionario, define una secuencia de covarianzas, $\sigma(0), \sigma(1), \ldots$, con función de autocorrelaciones

$$\rho_h = \sigma(h)/\sigma(0),\ h = \ldots,-1, 0, 1, \ldots.$$

la transformada de Fourier de la función de autocorrelaciones nos da la contribución de cada frecuencia entre $-\pi$ y π a la variación de y_t, la función de densidad espectral estandarizada es

$$f(\lambda) = \frac{1}{2\pi} + \frac{1}{\pi}\sum_{h=1}^{\infty}\rho_h \cos\lambda h$$
$$= \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\rho_h \cos\lambda h \qquad -\pi \le \lambda \le \pi.$$

Si $\sum_{h=-\infty}^{\infty}|\sigma(h)| < \infty$, entonces un modelo que representa una descomposición de la función de autocorrelación en componentes correspondientes a distintas frecuencias es

$$\rho_k = \int_{-\pi}^{\pi} f(\lambda)\cos\lambda k d\lambda \qquad k = \ldots, -1, 0, 1, \ldots$$

La información a través de la función de densidad espectral estandarizada $f(\lambda)$ es equivalente que la información a través de ρ_k, son lo que se llama un par de transformadas de fourier.

Una condición necesaria y suficiente para que exista ρ_k es que exista una función $F(\lambda)$ no decreciente, $F(\pi) = 1$, la función de distribución espectral, dado que $f(\lambda) = f(-\lambda)$ es

$$F(\lambda) = \int_{-\pi}^{\lambda} f(v)dv = 2\int_0^{\lambda} f(v)dv = 2\int_0^{\lambda}\left(\frac{1}{2\pi} + \frac{1}{\pi}\sum_{h=1}^{\infty}\rho_h \cos vh\right)dv = \frac{\lambda}{\pi} + \sum_{h=1}^{\infty}\rho_h \frac{sen\lambda h}{h}$$

donde $0 \le \lambda \le \pi$. Si $y_1, \ldots, y_T$ son T observaciones consecutivas del proceso estocástico $\{y_t\}$, $t = \ldots,-1, 0, 1, \ldots$, $E[y_t] = \mu$, $\sigma(h) = E[(y_t - \mu)(y_{t+h} - \mu)]$. Un estimador insesgado de $\sigma(h)$ es

$$C_h = C_{-h} = \frac{1}{T-h}\sum_{t=1}^{T-h}(y_t - \mu)(y_{t+h} - \mu) \qquad h = 0, 1, \ldots, T-1$$

ó

$$c_h = c_{-h} = (T - |h|)C_h/T \qquad h = 0, 1, \ldots, T-1$$

con sesgo $-|h|\sigma(h)/T$, y para algún h el límite de este es 0 si $T \to \infty$, luego asintóticamente insesgado.

Si la media del proceso no es conocida entonces podemos utilizar $\bar{y}$, la media de las T observaciones. La secuencia de autocorrelaciones muestrales es entonces definida como

$$r_h = r_{-h} = c_h / c_0, \quad h = -(T-1),\ldots,(T-1).$$

En esta situación definimos el espectrograma muestral

$$R^2(\lambda) = A^2(\lambda) + B^2(\lambda),$$

donde

$$A(\lambda) = \frac{2}{T}\sum_{t=1}^{T}(y_t - \mu)\cos\lambda t$$

y

$$B(\lambda) = \frac{2}{T}\sum_{t=1}^{T}(y_t - \mu)sen\lambda t.$$

$R(\lambda)$ es proporcional al coeficiente de correlación entre la serie observada y $\{\cos\lambda t, \text{sen}\,\lambda t\}$ una función trigonométrica de frecuencia $\lambda/2\pi$. El periodograma o sea la función de densidad espectral muestral, reemplazando ρ_h por r_h y si $\mu = 0$ es

$$I_T(\lambda) = \frac{T}{8\pi c_0}\frac{Tc_0}{Tc_0}R^2(\lambda) = \frac{1}{2\pi c_0}\left|\sum_{t=1}^{T} y_t e^{i\lambda t}\right|^2 = \frac{1}{2\pi}\sum_{h=-(T-1)}^{(T-1)} r_h \cos\lambda h \qquad -\pi \le \lambda \le \pi.$$

La función de distribución espectral muestral será

$$\hat{F}_T(\lambda) = \int_{-\pi}^{\lambda} I_T(v)dv = 2\int_0^{\lambda} I_T(v)dv = \frac{\lambda}{\pi} + 2\sum_{h=1}^{T-1} r_h \frac{sen\lambda h}{h} \qquad 0 \le \lambda \le \pi.$$

Siendo la función de densidad espectral muestral y la función de distribución espectral muestral, ambas, funciones lineales de las autocorrelaciones, notar que $f(\lambda)$ es una serie infinita con $\sigma(r), r = 0,1,\ldots,$ como coeficientes, $I(\lambda)$ no debería ser un estimador insesgado de $f(\lambda)$ ya que solo utiliza T de las correlaciones en esta situación $E[I(\lambda)] \neq f(\lambda)$.

Sin embargo si la serie

$$\sum_{r=-\infty}^{\infty}\sigma(r)\cos\lambda r$$

converge, entonces será un estimador asintóticamente insesgado de $f(\lambda)$, por tanto

$$\lim_{T\to\infty} E[I(\lambda)] = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} r_h \cos\lambda r$$

$$\lim_{T\to\infty} E[I(\lambda)] = f(\lambda).$$

2.3 Test de bondad de ajuste y el test estadístico KPSS

La distribución muestral asociada a una muestra aleatoria $y_1,\ldots,y_T$ puede utilizarse como aproximación de la distribución teórica $F(y)$ a partir de las observaciones, la distribución muestral es

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{T} I_{(-\infty,y]} y_i$$

y puede usarse para contrastar la hipótesis nula de que, por ejemplo, $F(y) = F_0(y)$, donde $F_0(y)$ es una distribución totalmente especificada. Existen diferentes test basados en la diferencia entre la distribución muestral y la teórica, claro está que

$$F_n(y) \xrightarrow{p} F(y), \quad n\to\infty$$

y

$$\sup_{-\infty<x<\infty} |F_n(y) - F(y)| \to 0, \quad n\to\infty,$$

conocido como el teorema de Glivenko-Cantelli.

La distribución asintótica de esta diferencia, usando el teorema central del límite es

$$\sqrt{n}(F_n(y) - F(y)) \to N(0,\, p(1-p)), \quad n\to\infty$$

donde $F_n(y)$ es una proporción muestral, $p = F(y)$, bajo la convergencia en distribución y

$$\sqrt{n}(F_n(y) - F(y))$$

converge a un proceso estocástico gaussiano llamado puente browniano o brownian bridge en el caso especial en el que la verdadera distribución de la población sea uniforme.

Diferentes test , como el test de Anderson-Darling, el test de Kolmogorov-Smirnov con funcional asociado $D_n = \sup|F_n(y) - F_0(y)|$ o el Cramér-von Mises con funcional

$$W_n^2 = n\int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 dF_0(x),$$

están basados en esta diferencia. Estos funcionales son invariantes a transformaciones de la variable aleatoria, luego, si hacemos la transformación monótona $z = F_0(x)$, $F_0^{-1}(z) = x$,

z esta distribuida uniformemente en el intervalo $[0,1]$, el estadístico de Cramér-von Mises se simplifica y

$$W_n^2 = n\int_0^1 [F_n(z) - z]^2 dz.$$

Entonces $y_n(z) = \sqrt{n}\{F_n(z) - z\}$ tiende a un proceso gaussiano $y(z)$ cuando $n \to \infty$. La distribución asintótica de W_n^2 es

$$W^2 = \int_0^1 y^2(z)dz\,,$$

la media de $y(z)$ es cero, la función de covarianzas cuando todos los parámetros son conocidos es

$$\rho_0(s,t) = \min(s,t) - st,$$

la distribución de un puente browniano o brownian bridge.

De hecho si consideramos el modelo autorregresivo de primer orden y definimos la hipótesis nula de estacionariedad contra la alternativa de que existe una componente no estacionaria, bajo esta situación, el test bajo la hipótesis nula puede formularse como

$$\lambda = \frac{T^{-2}\sum_{i=1}^{T}\left[\sum_{t=1}^{i}\varepsilon_t\right]^2}{s^2} > c,$$

donde $s^2 = T^{-1}\sum_{t=1}^{T}(u_t - \bar{u})^2$ y c es el llamado valor crítico del test.

La distribución asintótica del test bajo la hipótesis nula puede encontrarse observando que las desviaciones en sumas parciales sobre la media convergen a un brownian bridge, o dicho de otra forma

$$\sigma^{-1}T^{-1/2}\sum_{s=1}^{[Tr]}\varepsilon_s \to B(r), \quad r \in [0,1],$$

donde $[Tr]$ es el más grande entero menor o igual a Tr y $B(r) = W(r) - rW(1)$. Entonces

$$\lambda \to \int_0^1 B(r)^2 dr\,,$$

que no es sino la distribución de Cramér-von Mises.

Kwiatkowski, Phillips, Schmidt y Shim proponen un test de estacionariedad de primer orden basado en el proceso

$$S_T(r) = \frac{1}{\hat{w}\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} (u_t - \bar{u}_T),$$

donde $r \in [0,1]$, $\bar{u}_T$ es la media muestral de $\{u_t\}_{t=1}^{T}$ y $\hat{w}$ es un estimador consistente de

$$w^2 = \lim_{T\to\infty} E\left[\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (u_t - \bar{u}_T)\right)^2\right].$$

Se puede considerar el funcional $h(S_T(r))$, donde $h(\cdot)$ es la métrica de Cramér-von Mises. Luego el test estadístico KPSS es dado por

$$KPSS = \frac{1}{(\hat{w}T)^2} \sum_{k=1}^{T} \left(\sum_{t=1}^{k} (u_t - \bar{u}_T)\right)^2$$

y bajo la hipótesis nula de estacionariedad

$$KPSS \xrightarrow{d} \int_0^1 k(\alpha)^2 \, d\alpha,$$

donde $k(\alpha) = W(\alpha) - \alpha W(1)$ es un puente browniano estándar. Un estimador consistente de

$$w = \lim_{T\to\infty} T^{-1} E\left(S_T^2\right)$$

cuando u_t satisface las condiciones de Phillips y Solo (1989) para un proceso lineal es

$$s^2(r) = T^{-1} \sum_{t=1}^{T} u_t^2 + 2T^{-1} \sum_{s=1}^{r} w(s,r) \sum_{t=s+1}^{T} u_t u_{t-s},$$

donde $w(s,r) = 1 - s/(r+1)$.

Si nos basamos en la métrica de Kolmogorov-Smirnov entonces Xiao (2001)

$$KS = \underset{1\le k\le n}{Max} \frac{1}{\hat{w}\sqrt{T}} \left| \sum_{t=1}^{k} \hat{u}_t - \frac{k}{T} \sum_{t=1}^{T} \hat{u}_t \right|.$$

Jong et al. (2007) proponen una versión robusta del test KPSS basada en el siguiente proceso

$$I_T(r) = \frac{1}{\hat{\sigma}\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} sign(u_t - m_T),$$

donde m_T es la mediana muestral de $\{u_t\}_{t=1}^{T}$, $\hat{\sigma}^2$ es un estimador consistente de

$$\sigma^2 = \lim_{T\to\infty} E\left[\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} sign(u_t - m_T)\right)^2\right]$$

y

$$sign(x)=\begin{cases}1 \;\; si\; x>0\\ 0 \;\; si\; x=0\\ -1\, si\; x<0\end{cases}.$$

Aplicando la métrica de Cramér-von Mises al proceso $I_T(r)$ obtenemos el test estadístico IKPSS

$$IKPSS=\frac{1}{(\hat{\sigma}T)^2}\sum_{k=1}^{T}\left(\sum_{t=1}^{k} sign(u_t-m_T)\right)^2,$$

que bajo la hipótesis nula de estacionariedad

$$IKPSS\xrightarrow{d}\int_0^1 k(\alpha)^2\,d\alpha,$$

donde $k(\alpha)=W(\alpha)-\alpha W(1)$ es un puente browniano estándar y por tanto tiene la misma distribución límite que el test estadístico KPSS.

2.4 Primeros cálculos

La diferencia entre la función de distribución espectral muestral, llamado periodograma, y la distribución espectral multiplicado por la raíz cuadrada del tamaño muestral es

$$\sqrt{T}\left[\hat{F}_T(\lambda)-F(\lambda)\right],\ 0\le\lambda\le\pi,$$

luego

$$\sqrt{T}\left[\hat{F}_T(\lambda)-F(\lambda)\right]=\frac{2}{\pi}\sum_{h=1}^{T-1}\frac{sen\lambda h}{h}\sqrt{T}(r_h-\rho_h)-\frac{2}{\pi}\sum_{h=T}^{\infty}\frac{sen\lambda h}{h}\sqrt{T}\rho_h$$

es un proceso estocástico en el intervalo $[0,\pi]$ que converge a un proceso gaussiano. Si suponemos un proceso lineal $y_t=\mu+\sum_{s=-\infty}^{\infty}\gamma_s v_{t-s}$ donde

$$\sum_{h=-\infty}^{\infty}|\gamma_h|<\infty,\ \sum_{h=-\infty}^{\infty}|h|\gamma_h^2<\infty$$

y $\{v_t\}$ variables aleatorias independientes e idénticamente distribuidas con $E[v_t]=0$ y varianza finita $E\left[v_t^2\right]=\sigma^2<\infty$. Entonces se puede encontrar la distribución conjunta de $\sqrt{T}(r_1-\rho_1),\ldots,\sqrt{T}(r_m-\rho_m)$, cuando $T\to\infty$, la distribución es $N(0,W)$, donde $W=(w_{gh})$ y

$$w_{gh} = \sum_{r=-\infty}^{\infty} \left(\rho_{r+g}\rho_{r+h} + \rho_{r-g}\rho_{r+h} - 2\rho_h\rho_r\rho_{r+g} - 2\rho_g\rho_r\rho_{r+h} + 2\rho_g\rho_h{\rho_r}^2 \right)$$

$$= \frac{4\pi}{\sigma^2(0)} \int_{-\pi}^{\pi} (\cos vh - \rho_h)(\cos vg - \rho_g) f^2(v) dv \, .$$

Por otro lado si hacemos la transformación monótona

$$u = G(\lambda)/G(\pi), \ \lambda = G^{-1}[G(\pi)u] = \lambda(u)$$

y

$$G(\lambda) = 2\int_0^{\lambda} f^2(v) dv,$$

entonces el proceso transformado es

$$Y_T(u) = \sqrt{T}\left\{ \hat{F}_T[\lambda(u)] - F[\lambda(u)] \right\} \quad 0 \le u \le 1$$

y, en esta situación la distribución límite de $Y_T(u)$ converge cuando $T \to \infty$ a un proceso gaussiano con función de covarianzas

$$\begin{array}{c} 4\pi G(\pi)\{\min(u,v) - uv + q(u)q(v)\}, \quad 0 \le u, v \le 1 \\ q(u) = u - F\left\{G^{-1}[G(\pi)u]\right\}, \end{array}$$

luego la función de covarianzas de $Y_T(u) / \left[2\sqrt{\pi G(\pi)}\right]$ converge a

$$K(u,v) = h(u,v) + q(u)q(v),$$

donde $h(u,v)$ es la función de covarianzas de un brownian bridge, si $\tilde{B}(u), 0 < u \le v$, un es un movimiento browniano estándar y, $B(u), 0 < u \le v$, un brownian bridge, la expansión de $\{B(u), u \in [0,1]\}$ puede ser obtenida a partir de los valores y vectores propios de su función de covarianzas la cual es

$$E[B(u)B(v)] = \min(u,v) - uv \quad 0 \le u, v \le 1,$$

luego, claro esta, que por el teorema de Mercer's, la función de covarianzas puede ser expresada como

$$h(u,v) = E[B(u)B(v)] = \min(u,v) - uv = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} f_j(u) f_j(v),$$

la serie converge uniformemente y absolutamente a $h(u,v)$, siendo λ_j el autovalor propio de $f_j(u)$, el autovector propio, de la ecuación integral homogénea

$$\lambda \int_0^1 k(u,v) f(u) du = f(v).$$

El puente browniano tiene la representación

$$B(u) = \sum_{j=1}^{\infty} \frac{1}{\sqrt{\lambda_j}} X_j f_j(u),$$

y como $\lambda_j = (\pi j)^2$ y $f_j(u) = \sqrt{2} sen\, j\pi u$ entonces

$$B(u) = \sum_{j=1}^{\infty} \frac{\sqrt{2}}{j\pi} sen(j\pi u) X_j \quad u \in [0,1],$$

donde $\{X_j\}_{j \in \mathbb{N}}$ son variables aleatorias idénticamente distribuidas $N(0,1)$. El proceso $B(u)$ es un proceso gaussiano y dado que las trayectorias del movimiento browniano son continuas

$$Y_T(u) / \left[2\sqrt{\pi G(\pi)}\right] \to B(u) + q(u)X,$$

donde X tiene distribución $N(0,1)$ y, la matriz de covarianzas de $B(u) + q(u)X$ es

$$K(u,v) = \min(u,v) - uv + q(u)q(v),$$

un brownian bridge más un termino aleatorio independiente. Por otro lado y utilizando la identidad de Parseval

$$\int_0^1 B^2(u)du = \sum_{j=1}^{\infty} \frac{X_j^2}{\lambda_j},$$

con función característica

$$E\left[e^{it\sum_{j=1}^{\infty} \frac{X_j^2}{\lambda_j}} \right] = \prod_{j=1}^{\infty} \left(1 - \frac{2it}{\lambda_j}\right)^{-1/2}.$$

La función característica tiene una expresión alternativa $1/\sqrt{D(2it)}$, donde $D(\lambda)$ es el determinante de Fredholm, el cual puede ser aproximado evaluando numéricamente los valores propios asociados, el cual tiene la expresión

$$D(\lambda) = \prod_{j=1}^{\infty} \left(1 - \frac{\lambda}{\lambda_j}\right).$$

El proceso $B(u) + q(u)X$, $q(u) = \sum_{j=1}^{\infty} \alpha_j f_j(u)$ tiene la representación

$$B(u) + q(u)X = \sum_{j=1}^{\infty} \left(\frac{X_j}{\sqrt{\lambda_j}} + \alpha_j X \right) f_j(u),$$

donde $\alpha_j = \int_0^1 q(u) f_j(u) du$.

2.5 Cramér-von Misses

Si consideramos la hipótesis nula

$$H_0 : F(\lambda) = F_0(\lambda),$$

donde la función de distribución espectral muestral sigue una función totalmente especificada, en esta situación el criterio de Cramér-von Mises para contrastar la hipótesis nula es

$$\frac{1}{4\pi G(\pi)} \int_0^1 Y_T^2(u)du = \frac{T}{2\pi G^2(\pi)} \int_0^\pi \left[\hat{F}_T(\lambda) - F_0(\lambda)\right]^2 f_0^2(\lambda)d\lambda,$$

que es una integral de una función cuadrática de un proceso estocástico gaussiano, con función de covarianzas dada, que puede expresarse como

$$\begin{aligned} S &= \int_0^1 [B(u) + q(u)X]^2 du \\ &= \int_0^1 \left[\sum_{j=1}^{\infty} \left(\frac{x_j}{\sqrt{\lambda_j}} + \alpha_j X \right) f_j(u) \right]^2 du \\ &= \sum_{j=1}^{\infty} \left(\frac{x_j}{\sqrt{\lambda_j}} + \alpha_j X \right)^2 \\ &= \sum_{j=1}^{\infty} Y_j^2, \end{aligned}$$

y es una serie infinita, que puede aproximarse por una serie finita

$$S_N = \sum_{j=1}^{N} Y_j^2$$

y

$$E[S - S_N] = \sum_{j=N+1}^{\infty} \left(\frac{1}{\lambda_j} + \alpha_j^2 \right),$$

cuando $N \to \infty$ la distribución de S_N se aproxima a la de S, así como sus funciones características.

La matriz de covarianzas de Y_N es $E[Y_N Y_N'] = \Lambda_N + \alpha_N \alpha_N'$, donde el subíndice representa el elemento j-ésimo, Λ_N es una matriz diagonal con $1/\lambda_j$ en el elemento j-ésimo, la función característica de S_N es

$$E\left[e^{itY_N' Y_N}\right] = \left| I_N - 2it(\Lambda_N + \alpha_N \alpha_N') \right|^{-1/2} = \prod_{j=1}^{N} \left(1 - 2it\phi_{jN}\right)^{-1/2},$$

ϕ_{jN} es el j-ésimo cero de

$$\left|\Lambda_N + \alpha_N \alpha' - \phi I_N\right| = \left|\Lambda_N - \phi I_N\right| \psi(\phi),$$

con

$$\psi(\phi) = 1 + \sum_{j=1}^{N} \frac{\alpha_j}{\frac{1}{\lambda_j} - \phi},$$

el determinante de Fredholm de S_N es

$$D_N(v) = \left|I_N - v(\Lambda_N + \alpha_N \alpha_N')\right| = \prod_{i=1}^{N}\left(1 - \frac{v}{\lambda_i}\right)\left(1 - v^2 \sum_{j=1}^{N} \frac{\alpha_j^2}{\lambda_j - v} - v\sum_{j=1}^{N} \alpha_j^2\right),$$

que converge a $D(v)$.

Puntos asintóticos del test pueden ser obtenidos encontrando una aproximación para los valores propios utilizando el método de los coeficientes de fourier y una primera aproximación a partir de los ceros, ϕ, de $\psi(\phi)$, valores de α_j pueden ser obtenidos por integración numérica a partir de

$$\alpha_j = \int_0^1 q(u) f_j(u) du = \frac{2\sqrt{2}}{G\pi} \int_0^{\pi} sen\left[j\pi u \frac{G(\lambda)}{G(\pi)}\right]\left[\frac{G(\lambda)}{G(\pi)} - F(\lambda)\right] f^2(\lambda) d\lambda.$$

La distribución de S puede ser aproximada por

$$T = \sum_{j=1}^{N} \frac{1}{\lambda_j} X_j^2 + c$$

$\{X_j\}_{j \in \mathrm{N}}$ son variables aleatorias idénticamente distribuidas $N(0,1)$.

Finalmente el cálculo del estadístico para el contraste, a partir de la función de covarianzas muestral obtenida de la muestra, puede ser aproximado por

$$\begin{aligned}
&\frac{T}{2\pi G^2(\pi)} \int_0^{\pi} \left[\frac{2}{\pi} \sum_{h=1}^{T-1} \frac{sen\lambda h}{h}(r_h - \rho_h)\right]^2 \left[\frac{1}{2\pi} \sum_{r=-\infty}^{\infty} \rho_r e^{i\lambda r}\right]^2 d\lambda \\
&= \frac{T}{2\pi G^2(\pi)} \sum_{g,h=1}^{T-1} \frac{(r_g - \rho_g)(r_h - \rho_h)}{gh} \cdot \sum_{r,s=-\infty}^{\infty} \rho_r \rho_s \int_{-\pi}^{\pi} sen\lambda g sen\lambda h e^{i\lambda(r-s)} d\lambda \\
&= \frac{T}{8\pi^4 G^2(\pi)} \sum_{r=-\infty}^{\infty} \left[\sum_{g=1}^{T-1} \frac{(r_g - \rho_g)(r_{r+g} - \rho_{r-g})}{g}\right]^2.
\end{aligned}$$

2.6 Test para un $AR(p)$

2.6.1 Introducción

Si la función de densidad espectral depende funcionalmente de un vector parámetros $\theta \in \Theta, \theta' = (\theta_1, \ldots, \theta_p)$ a través de la función de autocorrelación $\rho(\theta) = (\rho_1(\theta), \ldots, \rho_q(\theta))'$, la función de densidad espectral será $f(\lambda;\theta)$, en esta nueva situación

$$\sqrt{T}\left[\hat{F}_T(\lambda) - F(\lambda;\hat{\theta})\right] = \frac{2}{\pi}\sum_{h=1}^{T-1}\frac{sen\lambda h}{h}\sqrt{T}\left(r_h - \rho_h(\hat{\theta})\right) - \frac{2}{\pi}\sum_{h=T}^{\infty}\frac{sen\lambda h}{h}\sqrt{T}\rho_h(\hat{\theta})$$

Si utilizamos el método de linealización de Taylor, la expansión en series de Taylor de $F(\lambda;t)$ es

$$F(\lambda;t) = F(\lambda;\theta) + \frac{dF(\lambda;\theta)}{d\theta}(t-\theta) + O\left(|t-\theta|^2\right) \qquad t \to \theta$$

y

$$\sqrt{T}\left[\hat{F}_T(\lambda) - F(\lambda;\hat{\theta})\right] = \frac{2}{\pi}\sum_{h=1}^{T-1}\frac{sen\lambda h}{h}\sqrt{T}\{r_h - \rho_h(\theta)\} - \frac{dF(\lambda;\theta)}{d\theta}\sqrt{T}(\hat{\theta}-\theta) + \circ_p(1).$$

Por otro lado si asumimos que una estimación del vector de parámetros puede ser realizada a partir de la función de autocorrelaciones muestral, $r = (r_1, \ldots, r_q)'$, $\hat{\theta} = \theta(r)$, siendo un proceso estacionario

$$y_t = \sum_{s=0}^{\infty}\gamma_s u_{t-s}, t = \ldots, -1, 0, 1, \ldots,$$

donde $r \xrightarrow{p} \rho(\theta)$ y $\hat{\theta} = \theta(r) \xrightarrow{p} \theta(\rho(\theta)) = \theta$, entonces

$$\sqrt{T}\left[\hat{F}_T(\lambda) - F(\lambda;\hat{\theta})\right] = \frac{2}{\pi}\sum_{h=1}^{T-1}\frac{sen\lambda h}{h}\sqrt{T}\{r_h - \rho_h(\theta)\} - \frac{dF(\lambda;\theta(\rho))}{d\rho_j}\sqrt{T}\{r_j - \rho_j(\theta)\} + \circ_p(1).$$

2.6.2 Distribución del test

Sea $\{y_t\}$ un proceso autorregresivo de orden p estacionario

$$y_t + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} = u_t,$$

donde u_t es un ruido blanco.

En este caso el vector de parámetros $\theta \in \Theta, \theta' = (\theta_1, \ldots, \theta_p)$ es $\theta = \beta = (\beta_1, \ldots \beta_p)'$, a través de las ecuaciones de Yule-Walker se puede poner en función de ρ o viceversa, esto es

$$0 = \sigma(0)\sum_{j=0}^{p} \beta_j \rho_{h-j} \quad h = 1,2,\ldots$$

La función de densidad espectral para un proceso autorregresivo de orden p o $AR(p)$,el cual puede ser invertido, claro esta, que todas las raíces del polinomio característico están dentro del circulo unidad, es

$$y_t = \sum_{r=0}^{\infty} \delta_r u_{t-r},$$

un proceso de medias móviles infinito, la función de densidad espectral puede ser obtenida de

$$\left| \sum_{j=0}^{p} e^{i\lambda j} \right|^2 f(\lambda) = \frac{\sigma^2}{2\pi}.$$

En el caso para $\rho_1, \ldots, \rho_p$ función de $\beta_1, \ldots, \beta_p$

$$h(\lambda;\beta) = \frac{\sigma^2}{2\pi \left| \sum_{j=0}^{p} \beta_j e^{i\lambda j} \right|^2} = \frac{\sigma^2}{2\pi \sum_{j,k=0}^{p} \beta_j \beta_k e^{i\lambda(j-k)}}.$$

Una estimación de los coeficientes autorregresivos β puede ser obtenida resolviendo las p ecuaciones normales de Yule-Walker a través de la muestra, por tanto

$\sum_{j=1}^{p} r_{h-j} b_j = -r_h, \quad h = 1, \ldots, p.$ y $\rho_h(b) = r_h, \quad h = 1, \ldots, p., \quad b \xrightarrow{p} \beta$ entonces $b_h = \beta_h(r)$

$$\sqrt{T}\{F_T(\lambda) - F(\lambda;b)\} = \frac{2}{\pi}\sum_{h=1}^{T-1} \frac{sen\lambda h}{h} \sqrt{T}\{r_h - \rho_h(b)\} + \circ_p(1) = \frac{2}{\pi}\sum_{h=p+1}^{T-1} \frac{sen\lambda h}{h} \sqrt{T}\{r_h - \rho_h(b)\} + \circ_p(1),$$

y para expresarlo en términos de $r_h - \rho_h(r)$ habrá que calcular $dF(\lambda;\theta)/d\theta'$, en este caso con respecto a β.

Si $\sqrt{T}\left\{\hat{F}_T(\lambda)-F(\lambda/\hat{\rho})\right\}$ tiene distribución normal con media cero y función de covarianzas

$$\lim_{T\to\infty} E[W(\lambda)W(v)] = 4\pi\left[G\{\min(\lambda,v)|\rho\rangle\}-\frac{1}{2\pi}F(\lambda|\rho)F(v|\rho)-\frac{4\pi}{1+\beta'\rho}\alpha'(\lambda)R^{-1}\alpha(v)\right],$$

donde

$$\begin{aligned} W(\lambda) &= \frac{2}{\pi}\sum_{h=1}^{T-1}\frac{sen\lambda h}{h}v_h-\sum_{j=1}^{p}\frac{dF(\lambda|\rho)}{d\rho_j}v_j \\ &= \frac{2}{\pi}\sum_{h=1}^{T-1}\frac{sen\lambda h}{h}v_h-\frac{2}{\pi}\sum_{h=1}^{\infty}\sum_{j=1}^{p}\frac{sen\lambda h}{h}\frac{d_h\rho_h(\rho)}{d\rho_j}v_j, \end{aligned}$$

la distribución asintótica de $\sqrt{T}\left\{\hat{F}_T(\lambda)-F(\lambda/\hat{\rho})\right\}$ es obtenida a partir de la distribución de

$$W(\lambda)=A(\lambda)-B(\lambda)$$

donde

$$A(\lambda)=\frac{2}{\pi}\sum_{h=1}^{T-1}\frac{sen\lambda h}{h}v_h$$

$$B(\lambda)=\sum_{j=1}^{p}\frac{dF(\lambda|\rho)}{d\rho_j}v_j=\frac{2}{\pi}\sum_{h=1}^{\infty}\sum_{j=1}^{p}\frac{sen\lambda h}{h}\frac{d\rho_h(\rho)}{d\rho_j}v_j\ .$$

2.7 El criterio de Kolmogorov-Smirnov

Si consideramos la hipótesis nula

$$H_0: F(\lambda)=F_0(\lambda),$$

donde la función de densidad espectral muestral sigue una función totalmente especificada, en este contexto el criterio de Kolmogorov-Smirnov para contrastar la hipótesis nula es

$$\frac{1}{2\sqrt{\pi G(\pi)}}\sup_{0\le u\le 1}|Y_T(u)|=\sup_{0\le\lambda\le\pi}\frac{\sqrt{T}}{2\sqrt{\pi G(\pi)}}\left|\hat{F}_T(\lambda)-F_0(\lambda)\right|$$

para un α, $0<\alpha<1$, y cuando $T\to\infty$

$$P\left\{\sup_{0\le u\le 1}|B(u)+q_0(u)X|\le c\right\}=1-\alpha\,.$$

Si $d=\sup_{0\le u\le 1}|q_0(u)|$ entonces tenemos las siguientes desigualdades

$$\sup_{0\le u\le 1}|B(u)+q_0(u)X|\le\sup_{0\le u\le 1}|B(u)|+|X|d,$$

$$P\left\{\sup_{0\le u\le 1}|B(u)|+|X|d\le c\right\}\le P\left\{\sup_{0\le u\le 1}|B(u)+q_0(u)X|\le c\right\}\le P\left\{\sup_{0\le u\le 1}|B(u)|\le c\right\},$$

la primera parte es el producto de convolución de $\sup_{0\le u\le 1}|B(u)|$ y de $d|X|$ y puede ser utilizado como aproximación de la segunda parte, esta aproximación es

$$P\left\{\sup_{0\le u\le 1}|B(u)|+|X|d\le c\right\}=$$
$$=2\phi\left(\frac{w}{d}\right)-1+4\sum_{j=1}^{\infty}(-1)^j\frac{\exp\left[-2j^2w^2/\left(1+4d^2\right)j^2\right]}{\sqrt{1+4d^2j^2}}\times\left[\phi\left(\frac{w}{d\sqrt{1+4d^2j^2}}\right)-\phi\left(\frac{-4dj^2w}{\sqrt{1+4d^2j^2}}\right)\right].$$

La transformada de Laplace de la convolución sería el producto de las transformadas y la parte final de la desigualdad si

$$z=\sup_{0\le u\le 1}B(u),$$

donde $B(u)$ es un brownian bridge entonces

$$F_z(x)=1-e^{-2x^2}.$$

Similarmente si definimos

$$z'=\sup_{0\le u\le 1}|B(u)|,$$

distribución que también es conocida, es

$$F_{z'}(x)=1+2\sum_{j=1}^{\infty}(-1)^j e^{-2j^2x^2},$$

aunque hay una serie infinita esta converge rápidamente y es también una aproximación de

$$P\left\{\sup_{0\le u\le 1}|B(u)+q_0(u)X|\le c\right\}=1-\alpha.$$

Los extremos de $q_0(u)$ pueden ser encontrados haciendo cero la derivada

$$\frac{d}{d\lambda}\left[\frac{G(\lambda)}{G(\pi)}-F_0(\lambda)\right]=2f_0(\lambda)\left[\frac{f_0(\lambda)}{G_0(\pi)}-1\right]$$

luego $f_0(\lambda)=0$ o $f_0(\lambda)=G_0(\pi)=\frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\rho_h^2$.

3. Funciones de autocorrelación robustas

3.1 Introducción

Como hemos visto los test de bondad de ajuste para series temporales se basan en la secuencia de autocorrelaciones o bien dependen funcionalmente de la función de autocovarianzas o de la función autocorrelaciones. Lo mismo ocurre, con los test de raíces unitarias, como es el caso del test de Dickey Fuller de raíces unitarias o el test KPSS de estacionariedad.

En el dominio de la frecuencia, la función de distribución espectral muestral y la función de distribución espectral teórica son funciones lineales de la función de autocorrelaciones y, se pueden utilizar para construir test de bondad de ajuste, por tanto se puede comparar si la función de distribución espectral de un proceso estocástico sigue una totalmente especificada. Entre ellos, como se ha visto, el test de Cramér-von Mises o el test de Kolmogorov-Smirnov.

Bajo este contexto hay que señalar que el coeficiente de correlación de Pearson no es una medida robusta entre dos variables, la estimación puede verse afectada por un outlier, Wilcox (2004). Además de por outliers puede verse afectada por la no linealidad, muestras heterogéneas y restricciones de rango.

3.2 Función de influencia de la función de autocorrelación

El coeficiente de correlación de Pearson no es una medida robusta entre dos variables, la estimación puede verse afectada por un outlier, Wilcox (2004). Además de por outliers puede verse afectada por la no linealidad, muestras heterogéneas y restricciones de rango.

Medidas específicas de influencia para funciones de autocorrelación han sido sugeridas por Chernick, Downing y Pike (1982), Lattin (1983) y Li & Hui (1987). Sin embargo, todas han sido desarrolladas sin recurrir a una teoría apropiada y sin especificar valores críticos para declarar una observación como sobre influyente o no. Solo Chernick et al. (1982) menciona en cierta medida estos aspectos.

Una observación sobre influyente será aquella observación que sensiblemente determina el valor de la función de autocorrelación.

Si consideramos una estimación de covarianza $C_n(X) \in PDS(p)$, entonces tenemos que la clase de todas las matrices definidas positivas de orden $p \times p$ es equivariante si $C_n(AX + v) = AC_n(X)A^t \quad \forall v \in \Re^p$ y A es una matriz no singular. Luego debe ser para algún $v \in \Re^p$

$$\varepsilon^*(C_n, AX + v) = \varepsilon^*(C_n, X).$$

Una medida de la robustez del estimador de covarianza es el punto de ruptura para espacios muestrales finitos, definido por Donoho y Huber (1983), y es definido como la fracción más pequeña, m/n, de outliers que hacen la estimación no acotada. En el caso de la matriz de covarianzas es definido como la fracción más pequeña, m/n, de outliers que pueden hacer la estimación no acotada, en este caso, si el más pequeño autovalor propio , $\lambda_p(C_n)$, tiende a 0 ó si, $\lambda_1(C_n)$, el más grande autovalor propio tiende a ∞.

$$\varepsilon^*(C_n, X) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{Y_m} D(C_n(X), C_n(Y_m)) = \infty \right\}.$$

Donde el supremo es tomado sobre todas las colecciones anómalas Y_m y

$$D(A, B) = \max\left\{ \left| \lambda_1(A) - \lambda_1(B) \right|, \left| \lambda_p(A)^{-1} - \lambda_p(B)^{-1} \right| \right\},$$

donde $\lambda_1(A) \ge \ldots \lambda_p(A)$ son los autovalores propios ordenados obtenidos de la matriz A.

El máximo punto de ruptura fue calculado por Davies (1987) y, es definido cuando ninguno de los $p+1$ puntos están contenidos en un hiperplano de dimensión menor que

p y, si $n \geq p+1$, el punto de ruptura de un estimador afín equivariante C_n es al menos de $[(n-p+1)/2]$ puntos o más.

Si consideramos un estimador con funcional asociado T y F la distribución desconocida común de las variables que forman la muestra, se puede estimar $T(F)$[1] por $T(F_n)$, siendo F_n la bien conocida función de distribución empírica.

Si tomamos el funcional T en la distribución contaminada $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_x$, la función de influencia, que mide la influencia que tiene el dato anómalo x sobre el sesgo asintótico del estimador que tiene funcional asociado T, basada en Hampel's (1974), tiene la expresión

$$IF(X, T(F), F) = \lim_{\varepsilon \to 0} \frac{T[(1-\varepsilon)F + \varepsilon\sigma_x] - T(F)}{\varepsilon},$$

con lo que, da lugar, si tenemos en cuenta que $\rho(k)$ es la función de autocorrelaciones en el retardo k, para una serie estacionaria, $E(X_t) = \mu$ y $V(X_t) = \sigma^2$, si asumimos que $Z_t = (X_t - \mu_t)/\sigma$ para cada valor de t, entonces

$$IF(X, T(F), F) = IF(Z, \rho(k), H),$$

donde H es la distribución de (Z_t, Z_{t+k}).

Dado $U_{i,k,1} = \frac{1}{2}\left[\frac{Z_i + Z_{i+k}}{\sqrt{1+\rho(k)}} + \frac{Z_i - Z_{i+k}}{\sqrt{1+\rho(k)}}\right]$ y $U_{i,k,2} = \frac{1}{2}\left[\frac{Z_i + Z_{i+k}}{\sqrt{1+\rho(k)}} - \frac{Z_i - Z_{i+k}}{\sqrt{1+\rho(k)}}\right]$,

entonces tenemos que

$$\left[1-\rho^2(k)\right]^2 U_{i,k,1}U_{i,k,2} = z_i z_{i+k} - \frac{1}{2}\rho(k)\left(z^2{}_i + z^2{}_{i+k}\right).$$

Luego la Función de influencia, IF, en la función de autocorrelación teórica $\rho(k)$, de algún par de observaciones retardadas k veces es definida como la matriz $n \times L$, siendo L un número fijo de retardos, es

---

[1] Si $T(F)$ es un funcional y estamos ante el problema de regresión, entonces $T(F_X) = \int y dF_X(y)$. La expansión de von Mises del funcional T alrededor de F es

$$T(G) = T(F) + \int \psi_1(x)d(G-F) + 1/2\int \psi_2(x_1, x_2)d(G-F)^2 + \ldots = T(F) + \sum_{m=1}^{\infty}\int \psi_m(x_1, \ldots, x_k)d(G-F)^k$$

que puede ser interpretado como una expansión en series de Taylor evaluada en $s=1$ de $T((1-s)F + sG) = T\left(F + s(G-F)^2\right)$. La función $\psi(x)$ es conocida como la función de influencia de T.

$$IF\left((z_i, z_{i+k}), \rho(k)\right) = z_i z_{i+k} - \frac{1}{2}\rho(k)\left(z^2{}_i + z^2{}_{i+k}\right),$$

y la función de influencia basada en Chernick et al. (1982) es

$$IF\left((z_i, z_{i+k}), \rho(k), H_1\right) = \left[1-\rho(k)\right]U_{i,k,1}U_{i,k,2}.$$

Además si α_{kk} es la función de autocorrelación parcial (PACF) entonces

$$IF\left((z_i, z_{i+k}), \alpha(k), H_1\right) = \left[1-\alpha^2(k)\right]U_{i,k,1}U_{i,k,2}.$$

Cleveland (1972) calcula la función de autocorrelación inversa (IACF), es

$$\rho_i(k) = \frac{\gamma_*(k)}{\gamma_*(0)},$$

donde

$$\gamma_i(k) = \int_{-\pi}^{\pi} e^{ikw} f_*(w)dw \text{ y } f_*(w) = \{f(w)\}^{-1}$$

es la inversa de la función de densidad espectral. La relación entre IACF Y ACF viene dada por Olewuezi and Shangodoyin (2005) es

$$\gamma_*(k) = \begin{cases} \dfrac{2\pi}{\gamma_0}, & k = 0 \\ 4\pi\displaystyle\sum_{j=1}^{k}\left(\frac{1}{j}\right)\left(\frac{1}{\gamma_j}\right)\sin 2j\pi, & k \neq 0 \end{cases}$$

Por tanto dada la dualidad entre IACF y ACF, la función de influencia puede ser escrita como

$$IF\left((z_i, z_{i+k}), \rho_i(k), H_1\right) = \left[1-\rho_i(k)\right]U'_{i,k,1}U'_{i,k,2}$$

donde

$$U_{i,k,1} = \frac{1}{2}\left[\frac{Z_i + Z_{i+k}}{\sqrt{1+\rho_i(k)}} + \frac{Z_i - Z_{i+k}}{\sqrt{1+\rho_i(k)}}\right] \text{ y } U_{i,k,2} = \frac{1}{2}\left[\frac{Z_i + Z_{i+k}}{\sqrt{1+\rho_i(k)}} - \frac{Z_i - Z_{i+k}}{\sqrt{1+\rho_i(k)}}\right].$$

Si asumimos un proceso estocástico estacionario, sabemos que si es gaussiano, esta completamente determinado por la media y su función de autocorrelaciones, en esta situación $IF\left((z_i, z_{i+k}), \rho_i(k), H\right)$ tiene la distribución de una constante multiplicada por el producto de dos variables aleatorias normales $U'_{i,k,1}$ y $U'_{i,k,2}$.

3.3 Estimación altamente robusta de la función de autocorrelación

La estimación de covarianza entre dos variables puede basarse en un estimador de escala, en esta situación Huber, (1981) y Gnanadesikan, (1997) proponen

$$C(X,Y)=\frac{1}{4\alpha\beta}\left[\mathrm{var}(\alpha X+\beta Y)-\mathrm{var}(\alpha X-\beta Y)\right] \quad \forall \alpha,\beta\in\Re,$$

y dado que, para una serie temporal, X e Y representan la misma variable entonces debe ser $\alpha=\beta=1$.

Por su parte Rousseeuw y Croux (1992,1993) proponen como estimador de escala Q_n, que es definido como

$$Q_n = c\left\{\left|Z_i - Z_j\right|;\ i<j\right\}_{(k)},$$

donde

$$k=\mathrm{int}\left[\frac{\binom{n}{2}+2}{4}\right]+1,$$

y el factor c es elegido para conseguir consistencia del estimador, en el caso Gausiano $c=2.2191$.

En esta situación un estimador de la función de autocovarianza es, si extraemos las primeras $n-h$ observaciones para producir un vector u de tamaño $n-h$ y las $n-h$ últimas observaciones para producir otro vector v,

$$\hat{\gamma}_Q(h,x)=\frac{1}{4}\left[Q^2_{n-h}(u+v)-Q^2_{n-h}(u-v)\right].$$

Este estimador tiene punto de ruptura temporal del 25%, el máximo posible en el caso de funciones de autocovarianza. Por su parte Q_n tiene una alta eficiencia asintótica gaussiana del 82.27%. Y la función de autocorrelaciones es

$$\hat{\rho}_Q(h,x)=\frac{Q^2_{n-h}(u+v)-Q^2_{n-h}(u-v)}{Q^2_{n-h}(u+v)+Q^2_{n-h}(u-v)},$$

siendo de esta forma que $\left|\hat{\rho}_Q(h,x)\right|\le 1$.

El punto de ruptura muestral de un estimador de escala $S_n(z)$ es

$$\varepsilon^*(S_n(z)) = \max\left\{\frac{m}{n} : \sup_Z S_n(z) < \infty \ y \ \inf_Z S_n(z) > 0\right\},$$

y el punto de ruptura de un estimador de covarianza basado en un estimador de escala S_n es

$$\varepsilon^*\left(\hat{C}_{S_n}(z)\right) = \max\left\{\frac{m}{n} : \sup_Z \hat{C}_{S_n}(z) < \infty \ y \ \inf_Z \hat{C}_{S_n}(z) > -\infty \ y \right.$$
$$\left. \inf_Z S_n\left(z(\alpha\beta)'\right) > 0 \ y \ \inf_Z S_n\left(z(\alpha-\beta)'\right) > 0\right\},$$

por su parte para una serie temporal tenemos que $u = (X_1,\ldots,X_{n-h})^t$ y $v = (X_{h+1},\ldots,X_n)^t$, entonces para $z = (u,v)$ es $\varepsilon_{n-h}{}^*\left(\hat{C}_{S_{n-h}}(z)\right)$ y, por tanto $\varepsilon_n{}^*\left(\hat{\gamma}_{S_n}(h,x)\right)$ con $x = (X_1,\ldots,X_n)^t$.

Sin embargo esta definición de punto de ruptura pierde todo su significado si trabajamos con series de tiempo y funciones de autocorrelación, por un lado la perturbación no es igual si se encuentra al inicio o en medio, además el efecto dependerá notablemente de la distancia del retardo h. En esta situación es necesario introducir una nueva definición, el punto de ruptura temporal de un estimador de autocovarianza. Es definido como

$$\varepsilon^t\left(\hat{\gamma}(h,x)\right) = \max\left\{\frac{m}{n} : \sup_{I_m}\sup_X S_{n-h}(u+v) < \infty \ y \ \inf_{I_m}\inf_X S_{n-h}(u+v) > 0 \ y \right.$$
$$\left. \sup_{I_m}\sup_X S_{n-h}(u-v) < \infty \ y \ \inf_{I_m}\inf_X S_{n-h}(u-v) > 0\right\}.$$

La función de influencia es una función de 1 en $\Re$ dada por Genton y Ma (1999) es

$$IF((u,v),T,F) = \frac{1}{2\alpha\beta}\left[S(F_1)IF((\alpha u+\beta v),S,F_1) - S(F_2)IF((\alpha u-\beta v),S,F_2)\right],$$

T es un estimador de γ, S es un estimador de escala, con función de influencia $IF(.,S,F_i)$ $i = 1,2$ y, asumimos las siguientes distribuciones $\alpha U+\beta V \approx F_1$, $\alpha U-\beta V \approx F_2$ y F es la distribución conjunta de U y V. En el caso de la función de autocovarianza, $\alpha = \beta = 1$, y bajo una distribución conjunta bivariada Gausiana F, la función de influencia de γ_Q es

$$IF((u,v),\gamma_Q,F) = \frac{1}{2}\left[\sigma^2_{U+V} IF\left(\frac{u+v}{\sigma_{U+V}},Q,\Phi\right) - \sigma^2_{U-V} IF\left(\frac{u-v}{\sigma_{U-V}},Q,\Phi\right)\right],$$

donde la función de influencia de Q_n en ϕ,Rousseeuw y Croux, (1993), es

$$IF(x,Q,\Phi)=c\frac{\frac{1}{4}-\Phi\left(x+\frac{1}{c}\right)+\phi\left(x-\frac{1}{c}\right)}{\int\phi\left(y+\frac{1}{c}\right)\phi(y)dy},$$

donde $c=2.2191$ y ϕ la distribución normal estándar.

Notar que la función de influencia de γ_Q está acotada entre

$$\pm\frac{\left[(a-b)\left(\sigma_U^2+\sigma_V^2\right)+2|a+b|\sigma_U\sigma_V\right]}{2},$$

donde $a=\max_x IF(x,Q,\Phi)$ y $b=\min_x IF(x,Q,\Phi)$.

Notar, además, que la definición es local en el sentido de que la definición de ruptura es solo para un vector retardado fijo.

3.4 Estimador de mínimo determinante de covarianza (MCD)

Los llamados M-estimadores son robustos en el sentido de que tienen función de influencia acotada pero su punto de ruptura es pequeño. Un estimador robusto con alto punto de ruptura para localización y escala es el estimador mínimo determinante de covarianza (MCD), *Minimum Covariance Determinant estimator*. En este estimador las estimaciones son dadas al minimizar el determinante de la matriz de covarianzas muestral clásica de $h=\left[(n+p+1)/2\right]$ datos de la muestra. Rousseeuw (1984), Butler, Davies y Jhun (1999) estudian la función de influencia y la eficiencia asintótica de MCD. Como característica, este método tiene mayor eficiencia que el método MVE, que estudiaremos posteriormente, y por tanto suele preferirse aunque tiene menor punto de ruptura.

El estimador MCD usa una función de pesos nula. Esto significa que la mitad de las observaciones consiguen un peso 1 y, la otra mitad 0 y pueden considerarse como outliers. Sin embargo, bajo este supuesto, puede ser difícil identificar outliers intermedios . Para ello la utilización de una función de pesos más general debería ser más apropiada y puede ser la propuesta por Ella Roelant, Stefan Van Aelst y Gert Willems obteniendo el estimador MWCD.

Así pues un modelo de localización y dispersión es dado por una distribución de $p\times 1$ vectores aleatorios x que esta completamente definido por un vector $p\times 1$ de

localización μ y una matriz simétrica definida positiva $p \times p$ de dispersión Σ. Las observaciones $x_i, i = 1, \ldots, n$ y es una matriz $n \times p$ W con n filas $x_1^t, \ldots, x_2^t$.

Dado $T(W)$ es un estimador multivariado de localización y dado $C(W)$ es un estimador de dispersión entonces la i-ésima distancia al cuadrado de Mahalanobis es el escalar

$$D_i^2 = D_i^2(T(W), C(W)) = (x_i - T(W))^t C^{-1}(W)(x_i - T(W))^t,$$

para cada observación x_i. Siendo para la distancia clásica de Mahalanobis como estimador de localización la media muestral y de dispersión la covarianza muestral.

Si consideramos ahora el subconjunto J_o de $C_n \approx n/2$ observaciones con determinante de la matriz de covarianza muestral más pequeña sobre todo $C(n, c_n)$ subconjuntos de tamaño c_n. Entonces dado T_{MCD} y C_{MCD} son las media muestral y la covarianza muestral de c_n casos en J_o. En esta situación el estimador de mínimo determinante de covarianza, $MCD(c_n)$, es $(T_{MCD}(W), C_{MCD}(W))$.

Para su calculo son necesarios el uso de algoritmos, uno de ellos se basa en el llamado conjunto elemental o remuestreo básico para estimadores robustos, este algoritmo usa K_n elementos de salida, subconjuntos aleatorios de $p+1$ casos, con p el número de variables. La j-ésima predicción elemental será el estimador clásico (T_j, C_j) calculados del j-ésimo conjunto elemental.

Otro algoritmo utiliza la técnica de concentración, en la cual no necesariamente el elemento de salida es elemental. De esta forma, dado $(T_{0,j}, C_{0,j})$ es la j-ésima salida y calculamos todas las n distancias de Mahalanobis $D_i(T_{0,j}, C_{0,j})$. En la siguiente iteración es calculado el estimador clásico $(T_{1,j}, C_{1,j})$ de las $C_n \approx n/2$ casos correspondiendo a las más pequeñas distancias. El proceso continua hasta por ejemplo k pasos y resulta la secuencia de estimadores $(T_{0,j}, C_{0,j}), (T_{1,j}, C_{1,j}), \ldots, (T_{k,j}, C_{k,j})$. El resultado de la iteración $(T_{k,j}, C_{k,j})$ es llamado el j-ésimo atractor. Tomando $k = 5$ pasos el algoritmo funciona bien y en el algoritmo de concentración el estimador final es el atractor que optimiza el criterio. El algoritmo de remuestreo básico es el caso especial con $k = 0$. Rousseeuw y Van Driessen (1999) prueban que el paso de concentración coje el determinante $|C_{i+1,j}| \leq |C_{i,j}|$ y

obtienen el llamado FMCD algoritmo de concentración. Tenemos básicamente que T rompe si d outliers pueden hacer que la medida

$$MED\left(\left\|w_i - T\left(W_d^n\right)\right\|\right),$$

sea arbitrariamente grande. Por lo tanto el estimador T no romperá si puede ser dibujado en una bola de radio R alrededor del origen.

Por su parte el estimador C rompe si el más pequeño autovalor propio λ_p tiende a 0 ó si λ_1 el más grande a ∞. Además es conocido que el mayor valor propio de $C_{p\times p}$ esta acotado superiormente por $p\max\left|c_{i,j}\right|$, con

$$c_{i,j} = \frac{1}{c_n - 1}\sum_{k=1}^{c_n}\left(z_{i,k} - \bar{z}_i\right)\left(z_{ji,k} - \bar{z}_j\right).$$

Por tanto λ_1 no puede ser arbitrariamente grande si z_i son todos contaminados en alguna bola de radio R alrededor del origen. Si todos los $\|z_i\|$ son acotados, entonces λ_i son acotados y, λ_p será 0 solo si el determinante de C es 0.

3.5 Estimador elipsoide de mínimo volumen (MVE)

El Estimador Elipsoide de Mínimo Volumen, (MVE), *Minimum Volume Ellipsoid estimator*, fue introducido por Rousseeuw (1985), es probablemente uno de los más conocidos (e.g. por He y Wang, 1996). Una de las razones es que alcanza un punto de ruptura casi igual a 0.5 frente a los M-estimadores multivariados de localización y escala, que basados en una muestra de una variable aleatoria p-dimensional, su punto de ruptura es a lo sumo $1/(1+p)$.

Este consiste en tomar como estimador de localización el centro del elipsoide más pequeño conteniendo al menos la mitad de los datos. Por tanto la estrategia consiste en encontrar la elipse con el área más pequeña que contiene la mitad de los datos. Así pues los puntos fuera de la elipse deberían considerarse como outliers.

Sin embargo este estimador no es $\sqrt{n}$ consistente, Davies (1992), haciendo menos atractiva su elección por razones de eficiencia. Otro estimador el cual tiene tasa de convergencia normal es el estimador MCD, Rousseeuw (1985). Las estimaciones son dadas por la media y la matriz de covarianza calculada de la mitad de los datos los cuales

constituyen el más pequeño determinante de su matriz de covarianza. Rousseeuw y Van Driessen (1997) proponen un algoritmo para computar MCD, el cual es extremadamente rápido. Las propiedades teóricas de MCD han sido investigadas por Butler (1982) y Butler, Davies y Jhun (1993), pero la distribución asintótica sigue siendo desconocida. En el caso particular de una dimensión la función de influencia del estimador de escala, MCD, ha sido calculada por Croux y Rousseeuw, (1992).

Así pues, dado $X = \{x_1, \cdots, x_n\}$ y $n \geq p+1$ puntos. El estimador elipsoide de mínimo volumen estima $t_n \in \Re^p$ y $C_n \in PDS(p)$ que minimiza el determinante de C sujeto a

$$\#\left\{i : (x_i - t)^t C^{-1} (x_i - t)^t \leq c^2\right\} \geq \left[\frac{n+p+1}{2}\right],$$

donde $t_n \in \Re^p$ y $C_n \in PDS(p)$ determinan el centro y la estructura de covarianza del elipsoide de mínimo volumen cubriendo al menos $[(n+p+1)/2]$. La elección de c, una constante fija, no tiene influencia en el valor de t_n, sin embargo si en la magnitud de C. Este valor puede ser elegido de acuerdo a alguna función de distribución, si asumimos una distribución elíptica $P_{\mu,\Sigma}$ con densidad $|\Sigma|^{-1/2} f\left[\left\{(x-\mu)^t \Sigma^{-1}(x-\mu)\right\}^{1/2}\right]$ entonces una elección de c podría ser el valor el cual

$$P_{\mu,\Sigma}\left\{(x-\mu)^t \Sigma^{-1}(x-\mu)\right\} \leq c^2 = \int_{\|x\| \leq c} f(\|x\|) dx = 1/2 .$$

El punto de ruptura si $p = 1$, t_n es el punto medio del intervalo más pequeño cubriendo al menos $\lfloor n/2 \rfloor + 1$ puntos, y C_n es proporcional al tamaño de ese intervalo, luego se necesita reemplazar al menos $[(n+1)/2]$ puntos para hacer $\|t_n\|$ infinitamente grande y reemplazando $(n/2)$ puntos en uno de los $n - [(n-1)/2]$ puntos para hacer C_n igual a 0.

Si $p \geq 2$ entonces $\varepsilon^*(t_n, X)$ y $\varepsilon^*(C_n, X)$ son al menos $[(n-p+1)/2]/n$.

Además el estimador Elipsoide de Mínimo Volumen puede considerarse un S-estimador, dado $X = \{x_1, \cdots, x_n\}$ y $n \geq p+1$ puntos. El estimador elipsoide de mínimo volumen estima $t_n \in \Re^p$ y $C_n \in PDS(p)$ que minimiza el determinante de C sujeto a

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left[\left\{(x_i - t)^t C^{-1}(x_i - t)^t\right\}^{1/2}\right] \leq b ,$$

donde $t_n \in \Re^p$ y $C_n \in PDS(p)$ determinan el centro y la estructura de covarianza del elipsoide. Las estimaciones MVE se obtienen cuando $nb = n - [(n+p+1)/2]$ y $\rho(\cdot) = 1 - _{[-c,c]}(\cdot)$.

Si ρ es simétrica, dos veces diferenciable, $\rho(0) = 0$ y además si existe una constante $c > 0$ tal que ρ es estrictamente creciente en $[0,c]$ y, constante en $[c,\infty]$, entonces si tomamos $r = b/\sup\rho$ y si $r \le (n-p)/2n$ el punto de ruptura es

$$\varepsilon^*(t_n, X) = \varepsilon^*(C_n, X) = [nr/n].$$

3.6 Estimador de correlación media biponderada

El estimador de correlación media biponderada, *Biweight Midcovariance estimator*, Wilcox (2004), es calculado de la misma manera que el coeficiente de correlación de Pearson, la diferencia se encuentra en las medidas para la media, desviaciones sobre la media y de covarianza utilizadas. El estimador de correlación media biponderada utiliza medidas robustas para su calculo.

Dado $(X_1,Y_1),\ldots,(X_n,Y_n)$ es una variable aleatoria de una distribución bivariada entonces, Wilcox (1997), dado

$$U_i = \frac{X_i - M_X}{9MAD_X}$$

donde M_X es la mediana, y MAD_X es la mediana de las desviaciones absolutas, entonces es

$$MAD_X = MEDIAN\{|X_1 - M_X|,\ldots,|X_i - M_X|\}.$$

Notamos que el punto de ruptura para tamaños muestrales finitos de MAD es de aproximadamente 0.5.

Dado V_i es una funcion de Y_i definida como U_i para X_i. El estimador entre X e Y es dado por

$$s_{bxy} = \frac{n\sum a_i (X_i - M_X)(1-U_i^2)^2 b_i (Y_i - M_Y)(1-V_i^2)^2}{\left(\sum a_i (1-U_i^2)(1-5U_i^2)\right)\left(\sum b_i (1-V_i^2)(1-5V_i^2)\right)},$$

donde

$$\begin{array}{llll} a_i = 1 & si\ -1 \le U_i \le 1, & si\ no\ a_i = 0 \\ b_i = 1 & si\ -1 \le V_i \le 1, & si\ no\ b_i = 0 \end{array}.$$

Una estimación de la correlación media biponderada entre X e Y es dada por

$$r_b = \frac{s_{bxy}}{\sqrt{s_{bxx} s_{byy}}},$$

donde s_{bxx} y s_{byy} son las estimación de varianza media biponderada de X e Y.

3.7 Estimador de correlación de porcentaje ajustado

Al igual que el estimador de correlación media biponderada, el propósito del estimador de correlación de porcentaje ajustado poblacional, *Population Percentage Bend Correlation estimator*, es el de analizar la correlación poblacional utilizando para ello medidas robustas. Como son la mediana y una generación de MAD, de esta forma el coeficiente de correlación Winsorizado es fijado antes más bien que cuando son determinados los datos.

El coeficiente de correlación $\alpha - Winsorizado$ poblacional es

$$\rho_W = \frac{E_W\left[\left(X - \mu_{\alpha,x}^W\right)\left(Y - \mu_{\alpha,y}^W\right)\right]}{\sigma_{W,x}\sigma_{W,y}},$$

que no es más que el coeficiente de correlación de la distribución Winsorizada bidimensional, basado en las medias $\alpha - Winsorizadas$ poblacionales y las varianzas $\alpha - Winsorizadas$.

Así pues el coeficiente de correlación de porcentaje ajustado, dadas las medianas poblacionales definidas como M_X y M_Y. Y si consideramos la función ψ_1 asociada al estimador de Huber con constante bend igual a $b = 1$,

$$\psi(x) = \max\{-1, \min(1, x)\},$$

con parámetros de escala w_X y w_Y, para un porcentaje de ajuste β $(0 \le \beta \le 0.5)$ es definido por las ecuaciones

$$P\{|X - M_X| < w_X\} = 1 - \beta,$$

$$P\{|Y - M_Y| < w_Y\} = 1 - \beta,$$

y los parámetros de localización η_X y η_Y tales que

$$E\left[\psi\left(\frac{X - \eta_X}{w_X}\right)\right] = 0,$$

$$E\left[\psi\left(\frac{Y-\eta_Y}{w_Y}\right)\right]=0\,.$$

Si definimos

$$U=\frac{X-\eta_X}{w_X}\,,$$

y

$$U=\frac{Y-\eta_Y}{w_Y}\,,$$

el coeficiente de correlación de porcentaje ajustado poblacional se define como

$$\rho=\frac{E[\psi(U)\psi(V)]}{\sqrt{E\left[\psi^2(U)\right]E\left[\psi^2(U)\right]}}\,.$$

Cuya estimación puede hacerse si definimos

$$R_i=\left|X_i-Me_x\right|,$$

donde Me_x es la mediana muestral y si m es igual a la parte entera de $(1-\beta)n$.

Definimos

$$\hat{w}_x=R_{(m)}\,,$$

y si i_1 es el número de X_i tales que

$$\frac{X_i-Me_x}{\hat{w}_x}<-1\,,$$

y si i_2 es el número de X_i tales que

$$\frac{X_i-Me_x}{\hat{w}_x}>1\,,$$

definimos

$$S_x=\sum_{j=i_1+1}^{n-i_2}X_{(j)}\ \mathrm{X},$$

y

$$\hat{\eta}_x=\frac{\hat{w}_x(i_2-i_1)+S_x}{n-i_1-i_2}\,,$$

entonces

$$U_i=\frac{X_i-\hat{\eta}_x}{\hat{w}_x}\,,$$

de igual forma si repetimos para Y_i

$$V_i = \frac{Y_i - \hat{\eta}_y}{\hat{w}_y},$$

finalmente si $A_i = \psi(U_i)$ y $B_i = \psi(V_i)$, se define el coeficiente de correlación de porcentaje ajustado muestral como

$$r = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}}.$$

4. Block Bootstrap

4.1 Introducción

Una serie de tiempo es una secuencia de sucesivas observaciones medidas sobre un conjunto de tiempos. La medidas de tiempo son usualmente discretas e igualmente espaciadas. Además las series de tiempo se caracterizan por existir dependencia entre las observaciones vecinas y frecuentemente la estructura de dependencia es muy compleja.

El método de Bootstrap, introducido por Efrom (1979), debe ser modificado para permitir la dependencia entre observaciones, fue primeramente propuesto para su uso en el dominio del tiempo por Freedman (1984) .

En las dos siguientes secciones se hace uso de un nuevo procedimiento, bastante sencillo, para generar muestras, por tanto es conveniente introducir el concepto de Block Bootstrap.

Block bootstrap consiste en dividir los datos en bloques de observaciones y muestrear los bloques aleatoriamente con reemplazamiento. Existen dos vertientes de remuestreo en el dominio del tiempo: remuestreo basado en modelos de Freedman (1984) y basado en bloques Hall (1995) y Carlstein (1986).

Por tanto en Block bootstrap se construyen muestras continuas de “bloques” de observaciones ordenadas y por remuestreo son añadidas y se añaden los bloques. Esta alternativa no paramétrica, block bootstrap, trata los bloques continuos de datos como unidades intercambiables para construir series remuestreadas.

Existen diferentes variantes de block bootstrap, entre ellas no sobrelapados, Hall (1985), Carlstein (1986) o sobrelapados, Hall (1985), Künsch (1989). La idea es que si hay

suficientes bloques, y si hay suficientes retardos entre las observaciones entre bloques, entonces mucha de la estructura de dependencia original estará en las muestras. Otros tipos son el block bootstrap circular propuesto por Politis y Romano (1992) y el bootstrap estacionario Politis y Romano (1994) en este último el tamaño del bloque es aleatorio.

Las dos variantes del tipo libre de modelo son Blockwise Bootstrap (BB) y Subseries approach (SA). En el BB, los bloques son remuestreados y añadidos para obtener una serie que, digamos, mimetiza la estructura de la serie original.

4.2 Block Bootstrap

La idea subyacente es la de construir bloques suficientemente grandes como para presentar suficiente estructura de dependencia de la serie original. Si elegimos un tamaño de bloque $l \approx n/k$, donde k es el número de bloques a remuestrear y para un proceso autorregresivo de orden uno, $AR(1)$, tenemos que

$$\rho_{block} = \left(\frac{k}{n}\right)0 + \left(\frac{n-k}{n}\right)\rho = \left(\frac{l-1}{l}\right)\rho .$$

Para construir los bloques consideremos el vector de observaciones consecutivas

$$Y_t = (X_{t-m+1}, \ldots, X_t), \quad t = m, \ldots, n,$$

entonces construir bloques solapados de vectores consecutivos

$$(Y_m, \ldots, Y_{m+l-1}), (Y_{m+1}, \ldots, Y_{m+l}), \ldots, (Y_{n-l+1}, \ldots, Y_n),$$

donde el parámetro $l \in \aleph$ es el tamaño del bloque. Si asumimos que el número de bloques $n-m+1 = kl$ con $k \in \aleph$, entonces el remuestreo de k bloques independientes con reemplazamiento es

$$Y_{S_1}, \ldots, Y_{S_1+l,}\ Y_{S_2}, \ldots, Y_{S_2+l}, \ldots, Y_{S_k}, \ldots, Y_{S_k+l,}$$

donde los puntos de inicio de los bloques $S_1, \ldots, S_k$ son variables aleatorias independientes e idénticamente distribuidas uniformemente en el intervalo $\{m-1, \ldots, n-l\}$. Si el número de bloques $n-m+1$ no es un múltiplo de l, entonces muestreamos $k = \lfloor (n-m+l)/l \rfloor + 1$ bloques pero usamos solo una porción del k ésimo bloque para conseguir $n-m+1$ muestras de m vectores. En esta situación para un estimador

$$\hat{\theta} = T\left(F_n^{(m)}\right),$$

donde

$$F_n^{(m)}(\cdot) = (n-m+1)^{-1}\sum_{t=m}^{n} I_{[Y_t \le \bullet]} ,$$

el estimador block bootstrap es

$$\hat{\theta}^{*B} = T\left(F_n^{(m)*B}\right),$$

$$F_n^{(m)*B}(\bullet) = (n+m-1)^{-1}\sum_{i=1}^{k}\sum_{t=S_i+1}^{S_i+l} I[Y_t \le \bullet].$$

Además si $\hat{\theta} = g_{n-m+1}(Y_m,\ldots,Y_n)$ es una función simétrica $g_{n-m+1}(\bullet)$ entonces

$$\hat{\theta}^{*B} = g_{n-m+1}\left(Y_{S_1+1},\ldots,Y_{S_1+l}, Y_{S_2+1},\ldots,Y_{S_2+l},\ldots,Y_{S_k+1},\ldots,Y_{S_k+l}\right).$$

4.3 Tamaño del bloque ℓ

El tamaño del bloque depende al menos de tres aspectos, el proceso que genera los datos, el estadístico a ser muestreado y la finalidad para la cual el bootstrap es usado, por ejemplo, estimación de sesgo, varianza de la función de distribución, etc.

Por tanto una de las cuestiones que se plantea es como conseguir un tamaño de bloque ℓ optimo. Para ello se pueden utilizar formulas explicitas del error cuadrático medio de estimación del tamaño del bloque. Una alternativa es usar Jackknife después del método Bootstrap, Hall et al. (1995), usar el método de remuestreo para construir un estimador del error cuadrático medio de estimación como función del tamaño del bloque y, entonces, minimizar este.

Carlstein (1986) y Künsch (1989) dan formulas asintóticas para el tamaño optimo de bloque bajo la regla de Carlstein y Künsch respectivamente. Así pues el tamaño optimo ℓ es proporcional a $n^{1/3}$ para estimación de sesgo y varianzas, sin embargo existe cierta incertidumbre de cual es la regla de proporcionalidad.

Hall, Horowitz y Jing (1995) muestran que el error cuadrático medio de la estimación $\hat{k}(n,l)$ donde $k = h(u)$ es proporcional a

$$\frac{1}{n^d}\left(\frac{C_1}{l^2} + \frac{C_2 l^c}{n}\right),$$

donde C_1 y C_2 dependen de k y de la estructura de covarianza, y $d=2$, $c=1$ si k es el sesgo o varianza, $d=1$, $c=2$ si k es el nivel de significación de un test de una cola, y $d=2$, $c=3$ si k es el nivel de significación en un test de dos colas.

Por tanto el tamaño optimo del bloque es igual a $(2C_1/C_2 n)^{1/3}$ para la estimación de la varianza y el sesgo de algún estimador $\hat{k}=h(\overline{X})$. Los coeficientes C_1 y C_2 no son óptimos bajo condiciones no asintóticas, esto es cuando el tamaño muestral es relativamente pequeño. También el calculo de C_1 y C_2 requiere estimaciones acuradas de $\gamma(h)$. Una solución propuesta por Hall, Horowitz y Jing (1995) es la de estimar empíricamente $MSE(\hat{k})$ como una función de l. Entonces dada una semilla inicial l_0 podemos encontrar una estimación de $\hat{k}(n,l_0)$. La serie es entonces dividida en subseries continuas de tamaño $m<n$ y para cada $j<l_0$, $\hat{k}(m,j)$ es calculada de $\{x_i,\ldots,x_{i+m-1}\}$ para $i=1,\ldots,n-m+1$. El error cuadrático medio es estimado por

$$MSE(m,k)=\frac{1}{n-m+1}\sum_{i=1}^{n-m+1}\left\{\hat{k}_i(m,j)-\hat{k}(n,l_0)^2\right\},$$

de esta forma si seleccionamos el tamaño del bloque $\hat{k}$ que minimiza el error cuadrático medio, entonces una estimación de $\hat{l}$ es dada por

$$\hat{l}=\hat{k}(n/m)^{1/(c+2)}.$$

5. Resultados empíricos

5.1 Introducción

En esta sección ilustramos el uso de experimento de Monte Carlo para calcular una estimación del tamaño, simplemente observando cuantas veces la hipótesis nula es rechazada en las muestras. El número de replicaciones ha sido de 2000, de esta forma podemos obtener una estimación de la potencia del test, esto es de la probabilidad de rechazar dado el valor del parámetro.

El modelo que se ha simulado es un proceso autorregresivo de orden 1 o AR(1), es

$$y_t = \mu + \beta t + u_t ; u_t = \phi u_{t-1} + \varepsilon_t ,$$

donde se ha tomado $\beta = 0, \mu = 0$ y ε_t son variables aleatorias independientes e idénticamente distribuidas $N\left(0, \sigma_\varepsilon^2\right)$. Además se ha considerado $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$ y un nivel de significación $\alpha \in \{0.01, 0.025, 0.05, 0.075, 0.1\}$. Los tamaños de muestra que se han utilizado son de 50, 100 y 200.

Los estimadores robustos, que se han utilizado para obtener una estimación robusta de la función de autocorrelaciones, para este caso de orden 1, $\rho(1)$, y que se han comparado, son junto con la estimación clásica, la estimación altamente robusta de Geman, MVE, MCD, la estimación de correlación de porcentaje ajustado y la estimación de covarianza media biponderada.

Para calcular la función de autocorrelaciones se ha extraído las primeras $n-h$ observaciones de $u=(u_1,\ldots,u_n)^t$ para crear v. De igual forma, extraer las últimas $n-h$ observaciones para crear w. Por tanto una estimación del parámetro puede ser dada por

$$\hat{\phi}_m = \hat{\rho}(1)\frac{mad(w)}{mad(v)}$$

donde $mad(x)=med_i\left(med_j\left(\left|x_i-x_j\right|\right)\right)$ y med_i es la mediana de las n medianas de $\left(\left|x_i-x_j\right|\right)$, $j=1,2,\ldots,n$.

Y el estadístico del test de Dickey-Fuller es

$$tau=(n)(\hat{\phi}-1).$$

Para estudiar la robustez de los estimadores se ha considerado un outlier aditivo de tamaño θ en cada una de las muestras, siendo t_0 una posición aleatoria en $1,\ldots,T$, el modelo es

$$x_{(t_0)}=u_{(t_0)}+\theta\sigma,$$

donde σ son variables aleatorias $N(0,\sigma_\varepsilon^2)$ y $\theta=10.000$.

Por otro lado se ha buscado la validez para generar muestras U_m perturbando el espacio muestral, creando así una variante dentro de la clase de algoritmos llamados block boostrap, para ello eliminar aleatoriamente z_1 observaciones iniciales, donde $z_1=k_1*m_1$ donde $k\approx U(0,1)$ y m_1 es elegido arbitrariamente. Y z_2 observaciones finales, donde $z_2=k_2*m_2$ donde $k_2\approx U(0,1)$ y m_2 es elegido arbitrariamente.

5.2 Resultados empíricos

Los resultados más significativos vienen dados en las tablas 5.1, 5.2, 5.3. Para una consulta detallada puede consultarse el anexo en donde se dan resultados para $\phi\in\{0.8, 0.85, 0.9, 0.95, 1\}$, $n\in\{50, 100, 200\}$ y en particular en las tablas 9.1, 9.2, 9.3 y 9.4 para un nivel de significación de $\alpha=0.01$. Si $\alpha=0.025$ se presentan los resultados en las tablas 9.5, 9.6, 9.7 y 9.8. Si $\alpha=0.05$ en las tablas 9.9, 9.10, 9.11 y 9.12. Para $\alpha=0.075$ en las tablas 9.13, 9.14, 9.15 y 9.16. Y finalmente si $\alpha=0.1$ en las tablas 9.17, 9.18, 9.19 y y 9.20. Para cada uno de los cinco grupos de tablas, la segunda tabla es obtenida

introduciendo un outlier aditivo, la tercera tabla perturbando el espacio muestral y la cuarta tabla perturbando e introduciendo un outlier aditivo.

Como se puede observar en las tablas 5.1, 5.2, 5.3, presentadas a continuación, bajo condiciones sin presencia de OA, todos los estimadores basados en métodos robustos son más potentes que el estimador clásico para tamaños de muestra de 50, 100 y 200, y para casi todos los niveles de significación. Sin embargo estos rechazarán la hipótesis de raíz unitaria en un mayor número de ocasiones que el estimador clásico.

En presencia de un OA los estimadores robustos prácticamente no se ven afectados en su potencia, si en cambio el estimador clásico que rechaza la presencia de una raíz unitaria en prácticamente casi el 100% de las simulaciones para cualquier tamaño de muestra y cualquier nivel de significación.

Dentro de los estimadores robustos destaca por su peor comportamiento el estimador de correlación media biponderado el cual presenta una menor potencia para un tamaño de muestra de 50. Además es el estimador que más veces rechaza la presencia de una raíz unitaria.

n=50	Sin OA						n=50	Con OA					
α=0,01			Population	Biweight			α=0,01			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,331	0,537	0,522	0,506	0,571	0,524	0,8	0,989	0,520	0,671	0,493	0,538	0,477
0,85	0,177	0,383	0,355	0,416	0,394	0,386	0,85	0,988	0,400	0,573	0,414	0,394	0,379
0,90	0,072	0,260	0,226	0,338	0,256	0,273	0,90	0,987	0,262	0,440	0,326	0,251	0,273
0,95	0,026	0,157	0,121	0,261	0,148	0,182	0,95	0,991	0,159	0,309	0,245	0,147	0,164
1	0,005	0,096	0,066	0,204	0,081	0,128	1	0,991	0,104	0,228	0,194	0,084	0,131
α=0,025			Population	Biweight			α=0,025			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,580	0,748	0,721	0,644	0,768	0,705	0,8	0,989	0,739	0,818	0,651	0,755	0,692
0,85	0,363	0,595	0,551	0,573	0,614	0,558	0,85	0,988	0,611	0,746	0,550	0,618	0,578
0,90	0,173	0,459	0,407	0,464	0,456	0,439	0,90	0,987	0,461	0,618	0,462	0,450	0,435
0,95	0,065	0,302	0,259	0,364	0,290	0,303	0,95	0,991	0,308	0,480	0,375	0,286	0,300
1	0,016	0,199	0,144	0,299	0,167	0,220	1	0,991	0,204	0,380	0,285	0,169	0,221
α=0,05			Population	Biweight			α=0,05			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,782	0,868	0,861	0,762	0,893	0,837	0,8	0,990	0,867	0,896	0,773	0,884	0,842
0,85	0,578	0,770	0,746	0,687	0,791	0,726	0,85	0,988	0,764	0,839	0,678	0,767	0,729
0,90	0,317	0,629	0,577	0,571	0,620	0,593	0,90	0,988	0,630	0,743	0,593	0,616	0,593
0,95	0,124	0,462	0,400	0,479	0,440	0,438	0,95	0,991	0,467	0,625	0,488	0,445	0,449
1,000	0,034	0,325	0,255	0,409	0,289	0,323	1,000	0,991	0,313	0,508	0,383	0,292	0,338
α=0,075			Population	Biweight			α=0,075			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,880	0,921	0,921	0,819	0,944	0,900	0,8	0,990	0,919	0,929	0,829	0,931	0,891
0,85	0,710	0,853	0,828	0,770	0,857	0,812	0,85	0,988	0,830	0,885	0,759	0,843	0,808
0,90	0,442	0,715	0,677	0,644	0,716	0,686	0,90	0,988	0,721	0,819	0,669	0,721	0,686
0,95	0,185	0,575	0,508	0,553	0,543	0,541	0,95	0,991	0,575	0,706	0,568	0,549	0,543
1,00	0,051	0,429	0,343	0,485	0,380	0,421	1,00	0,991	0,407	0,605	0,447	0,381	0,408
α=0,1			Population	Biweight			α=0,1			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,930	0,951	0,950	0,863	0,966	0,934	0,8	0,990	0,950	0,946	0,865	0,956	0,926
0,85	0,790	0,897	0,875	0,815	0,910	0,860	0,85	0,988	0,884	0,908	0,807	0,895	0,858
0,90	0,537	0,789	0,753	0,700	0,792	0,748	0,90	0,988	0,790	0,861	0,720	0,786	0,758
0,95	0,243	0,652	0,593	0,612	0,630	0,622	0,95	0,991	0,652	0,762	0,620	0,626	0,618
1	0,076	0,499	0,418	0,545	0,459	0,492	1	0,991	0,485	0,664	0,502	0,459	0,476

Tabla 5.1: Comparación de estimadores, $n = 50$, con OA y sin OA, $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$ y $\alpha \in \{0.01, 0.025, 0.05, 0.075, 0.1\}$.

n=100	Sin OA						n=100	Con OA					
α=0,01			Population	Biweight			α=0,01			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,936	0,947	0,949	0,885	0,967	0,914	0,8	0,993	0,949	0,971	0,881	0,969	0,917
0,85	0,697	0,819	0,812	0,745	0,852	0,777	0,85	0,995	0,823	0,896	0,764	0,849	0,786
0,90	0,317	0,536	0,500	0,541	0,559	0,518	0,90	0,996	0,559	0,705	0,554	0,580	0,535
0,95	0,069	0,260	0,221	0,325	0,251	0,281	0,95	0,994	0,257	0,444	0,317	0,256	0,266
1	0,009	0,098	0,076	0,180	0,088	0,139	1	0,993	0,091	0,225	0,158	0,088	0,115
α=0,025			Population	Biweight			α=0,025			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,989	0,988	0,990	0,951	0,992	0,974	0,8	0,993	0,991	0,990	0,948	0,995	0,984
0,85	0,888	0,935	0,925	0,880	0,946	0,903	0,85	0,995	0,934	0,957	0,888	0,942	0,910
0,90	0,563	0,756	0,711	0,705	0,752	0,711	0,90	0,996	0,767	0,844	0,715	0,777	0,720
0,95	0,167	0,449	0,387	0,477	0,438	0,437	0,95	0,994	0,440	0,622	0,465	0,441	0,436
1	0,023	0,196	0,153	0,292	0,174	0,224	1	0,993	0,186	0,368	0,258	0,173	0,206
α=0,05			Population	Biweight			α=0,05			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,997	0,997	0,996	0,984	0,998	0,993	0,8	0,993	0,999	0,991	0,985	1,000	0,994
0,85	0,967	0,974	0,969	0,944	0,977	0,959	0,85	0,995	0,977	0,978	0,954	0,981	0,961
0,90	0,762	0,876	0,843	0,818	0,879	0,837	0,90	0,996	0,893	0,917	0,839	0,902	0,862
0,95	0,293	0,626	0,560	0,620	0,609	0,595	0,95	0,994	0,625	0,752	0,614	0,610	0,587
1,000	0,043	0,325	0,248	0,395	0,285	0,339	1,000	0,995	0,301	0,506	0,368	0,273	0,311
α=0,075			Population	Biweight			α=0,075			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,999	0,999	0,999	0,991	1,000	0,996	0,8	0,993	1,000	0,992	0,995	1,000	0,998
0,85	0,986	0,986	0,982	0,969	0,989	0,977	0,85	0,995	0,992	0,984	0,972	0,993	0,980
0,90	0,868	0,929	0,909	0,876	0,940	0,899	0,90	0,996	0,942	0,943	0,903	0,942	0,912
0,95	0,417	0,730	0,664	0,697	0,717	0,682	0,95	0,994	0,736	0,817	0,699	0,720	0,687
1,00	0,065	0,424	0,338	0,475	0,384	0,428	1,00	0,995	0,401	0,594	0,452	0,361	0,388
α=0,1			Population	Biweight			α=0,1			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000	0,8	0,993	1,000	0,993	0,997	1,000	0,999
0,85	1,000	1,000	1,000	1,000	1,000	1,000	0,85	0,995	0,995	0,985	0,982	0,997	0,989
0,90	1,000	0,999	0,999	0,998	1,000	1,000	0,90	0,996	0,963	0,955	0,928	0,964	0,945
0,95	0,934	0,958	0,942	0,935	0,958	0,936	0,95	0,994	0,803	0,853	0,759	0,788	0,762
1	0,085	0,503	0,419	0,537	0,458	0,499	1	0,995	0,481	0,654	0,512	0,437	0,465

Tabla 5.2: Comparación de estimadores, $n = 100$, con OA y sin OA $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$ y $\alpha \in \{0.01, 0.025, 0.05, 0.075, 0.1\}$.

n=200	Sin OA						n=200	Con OA					
α=0,01			Population	Biweight			α=0,01			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	0,999	1,000	1,000	0,8	0,998	1,000	1,000	1,000	1,000	1,000
0,85	1,000	0,999	0,999	0,995	0,999	0,996	0,85	0,996	0,997	0,999	0,994	1,000	0,995
0,90	0,926	0,947	0,945	0,914	0,963	0,921	0,90	0,998	0,940	0,973	0,912	0,958	0,924
0,95	0,319	0,540	0,514	0,558	0,573	0,524	0,95	0,995	0,550	0,714	0,550	0,567	0,524
1	0,010	0,089	0,068	0,161	0,079	0,135	1	0,992	0,100	0,229	0,179	0,096	0,145
α=0,025			Population	Biweight			α=0,025			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000	0,8	0,998	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	0,999	0,998	1,000	1,000	0,85	0,996	1,000	0,999	1,000	1,000	0,999
0,90	0,987	0,988	0,990	0,974	0,994	0,975	0,90	0,998	0,985	0,991	0,968	0,991	0,978
0,95	0,565	0,758	0,713	0,728	0,767	0,720	0,95	0,996	0,753	0,846	0,724	0,762	0,713
1	0,021	0,194	0,147	0,274	0,174	0,229	1	0,992	0,202	0,385	0,285	0,185	0,245
α=0,05			Population	Biweight			α=0,05			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000	0,8	0,998	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	1,000	1,000	1,000	1,000	0,85	0,996	1,000	0,999	1,000	1,000	1,000
0,90	1,000	0,998	0,997	0,986	0,998	0,995	0,90	0,998	0,996	0,994	0,986	0,997	0,992
0,95	0,771	0,882	0,858	0,850	0,887	0,851	0,95	0,996	0,880	0,920	0,847	0,886	0,845
1,000	0,041	0,311	0,243	0,395	0,282	0,340	1,000	0,993	0,323	0,524	0,391	0,300	0,359
α=0,075			Population	Biweight			α=0,075			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000	0,8	0,998	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	1,000	1,000	1,000	1,000	0,85	0,996	1,000	0,999	1,000	1,000	1,000
0,90	1,000	0,999	0,999	0,993	0,999	0,999	0,90	0,998	0,998	0,995	0,994	0,999	0,996
0,95	0,877	0,930	0,913	0,908	0,936	0,905	0,95	0,996	0,928	0,942	0,898	0,930	0,906
1,00	0,061	0,416	0,341	0,478	0,382	0,433	1,00	0,994	0,433	0,607	0,478	0,394	0,443
α=0,1			Population	Biweight			α=0,1			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE	Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000	0,8	0,998	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	1,000	1,000	1,000	1,000	0,85	0,996	1,000	0,999	1,000	1,000	1,000
0,90	1,000	0,999	0,999	0,998	1,000	1,000	0,90	0,998	0,998	0,996	0,996	1,000	0,997
0,95	0,934	0,958	0,942	0,935	0,958	0,936	0,95	0,996	0,950	0,956	0,930	0,950	0,933
1	0,085	0,503	0,419	0,537	0,458	0,499	1	0,994	0,512	0,669	0,547	0,471	0,521

Tabla 5.3: Comparación de estimadores, $n = 200$, con OA y sin OA, $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$ y $\alpha \in \{0.01, 0.025, 0.05, 0.075, 0.1\}$.

6. Aprendizaje estadístico y el test de raíces unitarias

6.1 Introducción

El algoritmo que vamos a utilizar para contrastar raíces unitarias se basa en el algoritmo para el problema de clasificación llamado Random Forest, Breiman (2001), este utiliza una combinación lineal convexa de árboles de decisión en los cuales se modifican por un lado la distribución conjunta de la muestra de aprendizaje y por otro el algoritmo utilizado para construir los árboles. En el primer caso se puede construir una familia de transformaciones G_t en (Ω, A) en el espacio paramétrico Θ. Una transformación en el espacio muestral (χ, B) provoca una perturbación en P_θ y por tanto en una parte del espacio de medidas de probabilidad en (Ω, A). Así pues, cada perturbación dará lugar a una estimación diferente del parámetro en cada una de las muestras. Por otro lado se introduce, además, una perturbación en el algoritmo.

Así pues, con estas dos modificaciones estamos introduciendo aleatoriedad en la construcción de cada uno de los estimadores con lo que al combinar estos, deberá producirse una reducción en la estimación del riesgo esperado. La relación que existe entre el error del estimador combinado y el error de un estimador individual viene dada por

$$error_c = \frac{1+\ell(M-1)}{M} error + error_{bayes}$$

donde ℓ el coeficiente de correlación entre los errores de los estimadores[2] y M el número de estimadores. Si $\ell = 1$ el error del sistema combinado es igual al de un estimador simple. Si $\ell = 0$ el error del conjunto decrece proporcionalmente con el número de estimadores introducidos en la combinación.

6.2 Algoritmo para el calculo del test de raíces unitarias

6.2.1 Algoritmo

En este marco un algoritmo para calcular el test bajo la hipótesis nula de no estacionariedad y la presencia de una raíz unitaria puede construirse de la siguiente manera.

Dada una muestra de aprendizaje $z = \{(x_i)\}_{i=1}^{n}$.

1. Generar R estimadores de tamaño máximo M.

   Para $m = 1,\ldots, M * U(0,1)$.

   a. Generar una muestra U_m perturbando el espacio muestral, para ello eliminar aleatoriamente z_1 observaciones iniciales, donde $z_1 = k_1 * m_1$ donde $k \approx U(0,1)$ y m_1 es elegido arbitrariamente. Y z_2 observaciones finales, donde $z_2 = k_2 * m_2$ donde $k_2 \approx U(0,1)$ y m_2 es elegido arbitrariamente.

   b. Estimar una función de autocorrelaciones robusta, $\rho(h)$, con la muestra perturbada, eligiendo aleatoriamente una dentro de un conjunto de funciones robustas. Extraer las primeras $n-h$ observaciones de $u = (u_1,\ldots,u_n)^t$ para crear v. Extraer las últimas $n-h$ observaciones para crear w. La estimación del parámetro es

$$\hat{\phi}_m = \hat{\rho}(1)\frac{mad(w)}{mad(v)}$$

[2] la correlación entre $\hat{f}_i$ y $\hat{f}_j$ puede ser definida como la probabilidad de que $\hat{f}_i$ y $\hat{f}_j$ cometen el mismo error.

donde $mad(x) = med_i\left(med_j\left(\left|x_i - x_j\right|\right)\right)$ y med_i es la mediana de las n medianas de $\left(\left|x_i - x_j\right|\right)$, $j = 1,2,\ldots,n$

Calcular el estadístico del test de Dickey-Fuller

$$tau_m = (n - z_1 - z_2)(\hat{\phi} - 1)$$ y el p-valor asociado.

Y para un nivel de significación aceptar o rechazar.

2. Calcular el resultado del test para el R -ésimo estimador por agregación , por tanto hemos obtenido una combinación lineal convexa de los estimadores individuales. Este problema puede verse como un problema de clasificación de dos clases, en este caso aceptar o rechazar y por tanto se verifica

$$\hat{y} = \begin{cases} 1 & si \quad \sum_{i=1}^{m} I(tau_m \in Aceptación) > \sum_{i=1}^{m} I(tau_m \in \text{Re}chazo) \\ 2 & otro\ caso \end{cases}$$

o para el r-ésimo estimador mediante voto aleatorio, notar que esta alternativa puede verse como una imputación aleatoria, es

$$\hat{y} = 1 \cdot I\left(\frac{\sum_{i=1}^{n} I(y_i = aceptar)}{\sum_{i=1}^{n} I(x)} < d\right) + 2 \cdot I\left(\frac{\sum_{i=1}^{n} I(y_i = rechazar)}{\sum_{i=1}^{n} I(x)} \geq d\right)$$

donde $d \approx U[0,1]$.

En esta situación, no estamos más que utilizando la distribución de frecuencias de $y \in \{1,\ldots,J\}$. Así pues, construimos la distribución de frecuencia acumulada de y , seleccionamos un número aleatorio d de una distribución uniforme $[0,1]$ y entonces la predicción será

$$\hat{y} = j \quad \text{si} \quad d \in \left[\sum_{i=0}^{j-1} f_i, \sum_{i=1}^{j} f_i\right)$$

donde f_i es la frecuencia relativa de y , se cumple pues que $f_0 = 0$ y $\sum_{i=1}^{J} f_i = 1$.

3. Finalmente calcular el resultado del test mediante voto máximo del conjunto de R -estimadores. En caso necesario introducir una corrección para corregir el sesgo a favor de la hipótesis de no estacionariedad.

Tabla 6.1: Algoritmo para el calculo del test de raíces unitarias.

Así pues el problema puede verse como el de un clasificador f que pertenece a una familia $\Im$ de funcionales, en el problema de clasificación el conjunto de puntos $z_n = \{x_i, y_i\}_{i=1}^n$ es llamado pulverizable por $\Im$, si para todo vector binario $b \in \{-1,1\}^d$, que puede ser de aceptar la hipótesis nula o rechazar, existe un clasificador $f_b \in \Im$ tal que

$$\forall i \in \{1,\ldots,n\} \quad f_b(x_i) = b_i$$

Dicho de otra forma, si definimos la función

$$\Pi_{\Im}(n) = \max\left\{\left|\Im_{|Z}\right| : Z \subseteq X \times Y, |Z| = n\right\}$$

siendo $\Im_{|Z} = \{(x_1, y_1),\ldots,(x_n, y_n) : f \in \Im\}$, podemos elegir nuevamente $f : X \to \{-1,+1\}$. Claramente para todo n, si $\Im$ es finita $\Pi_{\Im}(n) \le |\Im|$ y $\Pi_{\Im}(n) \le 2^n$, resulta pues que $\Im$ pulveriza Z y por tanto el clasificador producirá todas las posibles clasificaciones de una muestra Z, entonces debe ser $\left|\Im_{|Z}\right| = 2^{|Z|}$.

Obviamente, el valor más grande de n para el que existe un conjunto z_n pulverizable es la llamada dimensión de Vapnik Chervonenkis, queda claro que para el caso que seguíamos estudiando, debe ser $|Y| = 2$ siendo $Y = \{-1,1\}$. Por tanto, la VC-dimensión del conjunto de clasificadores es definida como la cardinalidad del conjunto de puntos más grande X que puede ser clasificado por funciones en $\Im$, es decir

$$VC\dim(\Im) = \max\left\{|X| \,/\, \Im|X| = 2^{|Z|}\right\}$$

Si además suponemos que $\Im$ tiene VC-dimensión finita que podemos denotar por d y P es una medida de probabilidad arbitraria en $X \times Y$, en esta situación, con probabilidad al menos $1-\delta$, se verifica el resultado

$$R(f) \le \hat{R}(f) + \sqrt{\frac{1}{n}\left(d\left(1+\log\frac{2n}{d}\right)+\log\frac{\delta}{4}\right)}$$

Este resultado muestra que, si la VC-dimensión es finita, $VC\dim(\Im) = d < \infty$, entonces el riesgo muestral converge uniformemente a $R(f)$, lo que significa que para todo $\varepsilon > 0$ se tiene

$$\lim_{n\to\infty} P\left\{\sup_{f\in\Im}|R_n(f) - R(f)| > \varepsilon\right\} = 0$$

Este hecho puede considerarse una condición necesaria y suficiente para la consistencia del problema de minimización del riesgo empírico. Tiene que tenerse en cuenta que la condición de convergencia uniforme depende de $\Im$. Luego podría suponerse que, si tenemos en cuenta nuevamente que $VC\dim(\Im) = d < \infty$ y (χ, B) es un espacio de medida, la familia de funcionales $\Im$ en χ será llamada una clase Glivenko-Cantelli con respecto a una familia de medidas λ, con lo que para cada $\varepsilon > 0$ resulta que

$$\lim_{n\to\infty} \sup_{\mu\in\lambda} P\left\{ \sup_{m>n} \sup_{f\in\Im} \left| E_\mu f - E_{\mu_m} f \right| \geq \varepsilon \right\} = 0$$

siendo μ_n la medida empírica en las primeras n coordenadas de la muestra, este resultado probaría que f_n converge a f uniformemente en x con probabilidad uno.

De hecho, en caso de que la VC-dimensión fuera infinita, $VC\dim(\Im) = d = \infty$, entonces la minimización del riesgo empírico no podría asegurar la minimización del riesgo funcional y por tanto[3]

$$R(f) \leq \hat{R}(f) + \infty = \infty$$

Cabe señalar, en el primer resultado, el hecho de que la convergencia se produce independiente de la elección de P y que decrece con el tamaño de la muestra de aprendizaje, sin embargo la diferencia entre los errores aumenta con la complejidad de $\Im$. Dicho de otra forma, es inversamente proporcional al tamaño de la muestra y directamente proporcional a la VC-dimensión.

Por lo tanto, es necesario controlar la complejidad de $f \in \Im$. Un camino para ello es utilizar parámetros de regularización, si se verifica $\Im_1 \subset \Im_2 \subset \ldots \Im_s$ y $\Im_s = \{ f \in \Im \,/\, C(h) \leq s \}$ siendo $C(h)$ una medida de la complejidad, en esta situación y si

[3] $L(y, \hat{y})$ representa la penalización de escoger $\hat{y}$ cuando el verdadero valor es y, la pérdida media mide el riesgo teórico del funcional f

$$R_T(f) = E_Y[L(y, \hat{y})] = \int_{X\times Y} L(y, \hat{y}) dP(x, y) = \int_X \int_Y L(y, \hat{y}) dP(dy / x) P_X(dx)$$

El riesgo empírico o muestral es

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i), \ \{(x_i, y_i)\}_{i=1}^{n}, \quad \hat{R}_n(f) \overset{\wp}{\to} R(f), \ n \to \infty$$

utilizamos el criterio de minimizar el riesgo empírico, la decisión óptima será aquel valor de $f \in \Im$ para el cual se obtenga

$$\arg\min_{f\in\Im} \hat{R}(f,s) + \lambda C(h,s)$$

siendo λ un parámetro llamado de regularización. Este problema puede también ser planteado como

$$\min_{\substack{f\in\Im \\ s.a \\ C(h)\le S}} \hat{R}(f) \quad \text{o} \quad \min_{\substack{f\in\Im \\ s.a \\ \hat{R}(f)\le \alpha}} C(h)$$

en los tres casos puede corresponderse con un problema de optimización convexa.

Así pues, este problema puede ser definido como un problema de minimización del riesgo empírico regularizado, esto es

$$\hat{f}_{n,\lambda,s} = \arg\min_{f\in\Im} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda C(h,s)$$

que, desde luego, es una aproximación estocástica, si la $VC\dim(\Im)$ es finita, del problema de minimización del riesgo teórico regularizado, que viene dado por

$$f_{P,\lambda,s} = \arg\min_{f\in\Im} E_P[L(Y, f(X))] + \lambda C(h,s)$$

Para finalizar, cabe notar que la elección de un/os parámetro/s de regularización, así como una/s medida/s de la complejidad adecuada/s, será función de sus efectos al aplicarlos. De hecho, al controlar la complejidad del estimador, se puede estar incrementando el sesgo y a su vez reduciendo varianza. En esta situación, estamos ante el dilema clásico sesgo-varianza

6.2.2 Resultados empíricos

En esta sección ilustramos el uso de experimento de Monte Carlo para calcular una estimación del tamaño, simplemente observando cuantas veces la hipótesis nula es rechazada en las muestras. El número de replicaciones ha sido de 2000, de esta forma podemos obtener una estimación de la potencia del test, esto es de la probabilidad de rechazar dado el valor del parámetro. El modelo que se ha simulado es un proceso autorregresivo de orden 1,

$$y_t = \mu + \beta t + u_t;\ u_t = \phi u_{t-1} + \varepsilon_t,$$

donde $\beta = 0, \mu = 0$, ε_t son variables aleatorias independientes e idénticamente distribuidas $N\left(0, \sigma_\varepsilon^2\right)$ y se ha considerado $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$. El tamaño de muestra ha sido $n \in \{50, 100, 200\}$ y el nivel de significación $\alpha \in \{0.01, 0.025, 0.05, 0.075, 0.1\}$.

Los estimadores robustos, que se han utilizado para obtener una estimación robusta de la función de autocorrelaciones $\rho(1)$ son la estimación altamente robusta de Geman, MVE, MCD, la estimación de correlación de porcentaje ajustado y la estimación de covarianza media biponderada. El número de estimadores individuales utilizado ha sido de 500.

El procedimiento utilizado para obtener una estimación de la función de autocorrelaciones ha sido extrayendo las primeras $n-h$ observaciones de $u = (u_1, \ldots, u_n)^t$ para crear v. Extraer las últimas $n-h$ observaciones para crear w. Por tanto una estimación del parámetro puede ser

$$\hat{\phi}_m = \hat{\rho}(1)\frac{mad(w)}{mad(v)},$$

donde $mad(x) = med_i\left(med_j\left(\left|x_i - x_j\right|\right)\right)$ y med_i es la mediana de las n medianas de $\left(\left|x_i - x_j\right|\right)$, $j = 1, 2, \ldots, n$

Y el estadístico del test de Dickey-Fuller es

$$tau_m = (n - z_1 - z_2)\left(\hat{\phi} - 1\right).$$

Para estudiar la robustez de los estimadores se ha considerado un outlier aditivo de tamaño θ en cada una de las muestras, siendo t_0 una posición aleatoria en $1, \ldots, T$.

$$x_{(t_0)} = u_{(t_0)} + \theta\sigma,$$

donde σ son variables aleatorias $N\left(0, \sigma_\varepsilon^2\right)$ y $\theta = 10.000$.

Los resultados obtenidos vienen dados en las tablas 6.2 ,6.3, 6.4 y 6.5.

n=200				Φ		
α=0.05		0,80	0,85	0,90	0,95	1
Sin OA	Aprendizaje	1,000	1,000	0,996	0,801	0,187
	Clásico	1,000	1,000	1,000	0,771	0,041
	Higly(Coman)	1,000	1,000	0,998	0,882	0,311
	Population Percentage Bend (	1,000	1,000	0,997	0,858	0,243
	Biweight Midcovariance	1,000	1,000	0,986	0,850	0,395
	MCD	1,000	1,000	0,998	0,887	0,282
	MVE	1,000	1,000	0,995	0,851	0,340
Con OA	Aprendizaje	1,000	1,000	0,989	0,791	0,165
	Clásico	0,998	0,996	0,998	0,996	0,993
	Higly(Coman)	1,000	1,000	0,996	0,880	0,323
	Population Percentage Bend (	1,000	0,999	0,994	0,920	0,524
	Biweight Midcovariance	1,000	1,000	0,986	0,847	0,391
	MCD	1,000	1,000	0,997	0,886	0,300
	MVE	1,000	1,000	0,992	0,845	0,359

Tabla 6.2: Comparación de estimadores. Sin outlier aditivo y con outlier aditivo, $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}, \alpha \in \{0.05\}$. 500 estimadores individuales para el estimador basado en aprendizaje.

n=200	Sin OA						
α=0,01				Population	Biweight		
Φ	Aprendizaje	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,90	0,854	0,9255	0,947	0,945	0,914	0,963	0,921
0,95	0,362	0,319	0,540	0,514	0,558	0,573	0,524
1	0,033	0,0095	0,089	0,068	0,161	0,079	0,135
α=0,025				Population	Biweight		
Φ	Aprendizaje	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,90	0,951	0,987	0,988	0,990	0,974	0,994	0,975
0,95	0,592	0,565	0,758	0,713	0,728	0,767	0,720
1	0,087	0,0205	0,194	0,147	0,274	0,174	0,229
α=0,05				Population	Biweight		
Φ	Aprendizaje	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,90	0,989	0,9995	0,998	0,997	0,986	0,998	0,995
0,95	0,791	0,771	0,882	0,858	0,850	0,887	0,851
1,000	0,165	0,0405	0,311	0,243	0,395	0,282	0,340
α=0,075				Population	Biweight		
Φ	Aprendizaje	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,90	0,996	1	0,999	0,999	0,993	0,999	0,999
0,95	0,837	0,877	0,930	0,913	0,908	0,936	0,905
1,00	0,240	0,0605	0,416	0,341	0,478	0,382	0,433
α=0,1				Population	Biweight		
Φ	Aprendizaje	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,90	0,999	1	0,999	0,999	0,998	1,000	1,000
0,95	0,879	0,934	0,958	0,942	0,935	0,958	0,936
1	0,296	0,0845	0,503	0,419	0,537	0,458	0,499

Tabla 6.3: Comparación de estimadores. Sin outlier aditivo. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.9, 0.95, 1\}$, $\alpha \in \{0.01, 0.025, 0.05, 0.075, 0.1\}$. 500 estimadores individuales para el estimador basado en aprendizaje.

n=200	Con OA						
α=0,01				Population	Biweight		
Φ	Aprendizaje	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,90	0,844	0,998	0,940	0,973	0,912	0,958	0,924
0,95	0,365	0,995	0,550	0,714	0,550	0,567	0,524
1	0,056	0,9915	0,100	0,229	0,179	0,096	0,145
α=0,025				Population	Biweight		
Φ	Aprendizaje	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,90	0,954	0,998	0,985	0,991	0,968	0,991	0,978
0,95	0,592	0,9955	0,753	0,846	0,724	0,762	0,713
1	0,118	0,992	0,202	0,385	0,285	0,185	0,245
α=0,05				Population	Biweight		
Φ	Aprendizaje	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,90	0,996	0,998	0,996	0,994	0,986	0,997	0,992
0,95	0,801	0,996	0,880	0,920	0,847	0,886	0,845
1,000	0,187	0,993	0,323	0,524	0,391	0,300	0,359
α=0,075				Population	Biweight		
Φ	Aprendizaje	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,90	0,994	0,998	0,998	0,995	0,994	0,999	0,996
0,95	0,840	0,996	0,928	0,942	0,898	0,930	0,906
1,00	0,275	0,9935	0,433	0,607	0,478	0,394	0,443
α=0,1				Population	Biweight		
Φ	Aprendizaje	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,90	0,996	0,998	0,998	0,996	0,996	1,000	0,997
0,95	0,894	0,996	0,950	0,956	0,930	0,950	0,933
1	0,336	0,9935	0,512	0,669	0,547	0,471	0,521

Tabla 6.4: Comparación de estimadores. Con outlier aditivo. $\phi \in \{0.9, 0.95, 1\}$, $\alpha \in \{0.01, 0.025, 0.05, 0.075, 0.1\}$. 500 estimadores individuales para el estimador basado en aprendizaje.

Como se puede observar en las tablas 6.2, 6.3, sin presencia de OA, todos los estimadores basados en métodos robustos son más potentes que el estimador clásico. Sin embargo estos rechazarán la hipótesis de raíz unitaria en un número mayor de ocasiones que el estimador clásico. Con el estimador basado en aprendizaje se ha conseguido en cierta medida una reducción de este porcentaje.

En presencia de un OA, tablas 6.2, 6.4, los estimadores robustos prácticamente no ven afectada su potencia, si en cambio en el estimador clásico tal y como ya se ha comentado en el capitulo 5.

Como se aprecia, además, en la tabla 6.4 el estimador clásico rechaza en casi el 100% de las muestras, y para todos los niveles de significación, la presencia de una raíz unitaria bajo condiciones de un OA . El estimador bajo aprendizaje parece ser no tan potente como los estimadores individuales robustos aunque rechazará la presencia de una raíz unitaria en un porcentaje menor que estos.

α=0,01						
		Sin OA			Con OA	
Φ	50	100	200	50	100	200
0,95	0,068	0,151	0,322	0,064	0,112	0,375
1	0,052	0,040	0,044	0,032	0,052	0,064
α=0,025						
		Sin OA			Con OA	
Φ	50	100	200	50	100	200
0,95	0,147	0,311	0,570	0,124	0,251	0,598
1	0,100	0,112	0,106	0,072	0,116	0,108
α=0,05						
		Sin OA			Con OA	
Φ	50	100	200	50	100	200
0,95	0,251	0,434	0,753	0,231	0,418	0,761
1	0,163	0,175	0,183	0,131	0,199	0,212
α=0,075						
		Sin OA			Con OA	
Φ	50	100	200	50	100	200
0,95	0,351	0,526	0,829	0,331	0,530	0,841
1	0,231	0,279	0,227	0,212	0,291	0,299
α=0,1						
		Sin OA			Con OA	
Φ	50	100	200	50	100	200
0,95	0,426	0,618	0,875	0,390	0,606	0,900
1	0,291	0,339	0,299	0,299	0,343	0,359

Tabla 6.5: Comparación del estimador basado en aprendizaje en función del tamaño de muestra $n \in \{50,100,200\}$. Sin outlier aditivo y con outlier aditivo. $\phi \in \{0.95,1\}$, $\alpha \in \{0.01, 0.025, 0.05, 0.075, 0.1\}$. 500 estimadores individuales.

En primer lugar, hacemos notar que, en particular, para esta tabla, que las simulaciones se han realizado con 250 muestras, en el resto de tablas las simulaciones se han realizado con 2000 muestras.

Luego en función del tamaño muestral, para una simulación con 250 muestras, se observa la mayor potencia del test para muestras pequeñas que el test de Dickey-Fuller y se vuelven a corroborar todas conclusiones que se han extraído tanto en el capitulo 5, resultados dados en las tablas 5.1, 5.2 y 5.3, como en el capitulo 6, tablas 6.2, 6.3 y 6.4.

6.3 Fundamentos del algoritmo

6.3.1 Introducción

Para reducir la variabilidad en estimadores inestables, que se caracterizan por tener alta variabilidad, como veremos, se pueden utilizar técnicas de regularización, de perturbación o de agregación/combinación de estimadores.

Entre ellas se encuentra el algoritmo Bagging. Esta metodología utiliza tanto técnicas de perturbación como de combinación para reducir la variabilidad en estimadores inestables. Otro tipo de algoritmos que utilizan la agregación de una familia de modelos, son los considerados en base a una familia de modelos adaptativos, un caso es el algoritmo Boosting.

6.3.2 Perturbar el espacio muestral

Un camino para reducir la variabilidad es perturbando o modificando algún aspecto del problema como la función de perdida, la distribución de las variables o algún aspecto concreto del algoritmo. En este caso, seleccionar un funcional $f \in \Im$ el cual minimice el riesgo empírico para cada uno de los problemas modificados y utilizar estos para obtener una solución del problema. Claro tiene que estar ahora que esta solución puede obtenerse por agregación/combinación de cada uno de los problemas modificados.

De manera que podemos construir una familia de transformaciones G_t en (Ω, A) en el espacio de parámetros Θ, si $(\Omega, \mathrm{A}, \mu)$ es un espacio de medida completo de una familia de distribuciones de probabilidad $P_\theta = \{P_\theta / \theta \in \Theta\}$ en (Ω, A), una transformación en el espacio muestral (χ, B) provoca una perturbación en P_θ y por tanto en una parte del espacio de medidas de probabilidad en (Ω, A).

La técnica que podemos utilizar para obtener una familia de perturbaciones es modificando la distribución empírica $F_n = 1/n \sum_{i=1}^{n} \delta_{z_i}$, siendo δ_z la medida de Dirac que da masa 1 a z. Por lo tanto, si reponderamos aleatoriamente la distribución empírica se obtiene que $F_n^m = \sum_{i=1}^{n} w_{im} \delta_{z_i}$ para $m = 1, \ldots, M$, siendo $\{w_{im}\}_{i=1}^{n}$ una función de pesos la cual puede seguir alguna distribución de probabilidad con $E[w_{im}] = 1/n$. Más exactamente, si tomamos para los pesos $\{w_{im}\}_{i=1}^{n}$ una distribución multinomial, $w_{im} \in \{0,1,\ldots,n\}$ con

probabilidad $1/n$, en este caso cada perturbación se basa en una muestra Bootstrap[4] con reemplazamiento. Notar que si utilizamos una distribución hipergeométrica estaremos en el caso sin reemplazamiento.

Un camino alternativo es perturbar la función de perdida $L_m\left(Y,\hat{Y}\right)=L\left(Y,\hat{Y}\right)+\delta_m$ a través de alguna función aleatoria δ_m.

Con estas dos metodologías, perturbando el espacio muestral o la función de perdida, podemos obtener M estimaciones de los parámetros, en particular para el modelo de predicción lineal, el cual puede expresarse de la forma

$$f(x;a)=a_0+\int a(\theta)f(x;\theta)d\theta,\ \theta=(\theta_1,\ldots,\theta_k),\ \theta\in\Theta$$

la función de coeficientes óptimo debe ser

$$a^*=\arg\min_a E[L[y,\ f(x;a)]]$$

Ahora bien, si consideramos en una de las M muestras, teniendo en cuenta además que si f [5] es una función simple, se puede comprobar que

$$f(x;a)\approx a_0+\sum_{l=1}^{L}v_l a(\theta_l)f(x;\theta_l)=c_0+\sum_{l=1}^{L}c_l f(x;\theta_l)$$

resulta pues que, los coeficientes estimados para el modelo usando esta muestra y si minimizamos el riesgo empírico regularizado son

$$\{\hat{c}_l\}_{l=1}^{L}=\arg\min_{\{\hat{c}_l\}_{l=1}^{L}}\frac{1}{n}\sum_{i=1}^{n}L\left[y_i,c_0+\sum_{l=1}^{L}c_l f(x_i;\theta_l)\right]+\lambda\sum_{l=1}^{L}h\left(\left|c_l-c_{l0}\right|\right)$$

siendo λ un parámetro de regularización y la función $h(\cdot)$ puede ser elegida tal que

$$h(x)=x^{\delta},\ \delta\geq 0$$

Notar que diferentes elecciones de λ y δ pueden tener como resultado un incremento en la acuracidad.

Llegados a este punto para las M muestras, podemos suponer Bootstrap, entonces si $S_m\subset\{x_i,y_i\}_{i=1}^{n},\ m=1,\ldots,M$ se tiene

---

[4] Una muestra Bootstrap de tamaño n se construye tomando una muestra aleatoria simple del espacio muestral.

[5] Si (Ω,A,μ) es un espacio de medida. Una función $f:\Omega\to\Re$ se dice simple, si existe una partición $\{E_i:1\leq i\leq n\}$ de Ω por elementos de A y un conjunto $\{a_i:1\leq i\leq n\}$ de números reales, con $a_i\neq a_j$ si $i\neq j$ tal que $f=\sum_{i=1}^{n}a_i\chi_{E_i}$ donde χ_{E_i} denota la función característica de E_i respecto Ω. Para este tipo de funciones se define la integral de f respecto a μ como $\int_\Omega f d\mu=\sum_{i=1}^{n}a_i\mu(E_i)$.

$$\{\hat{c}_{lm}\}_{l=1}^{L} = \underset{\{\hat{c}_{ml}\}_{l=1}^{L}}{\arg\min} \frac{1}{n} \sum_{i \in S_m} L\left[y_i, c_{0m} + \sum_{l=1}^{L} c_{lm} f\left(x_i; \theta_{lm}\right) \right] + \lambda \sum_{l=1}^{L} h\left(\left| c_{lm} - c_{l0m} \right|\right),$$

$$m = 1, \ldots, M$$

La estimación final puede ser una combinación lineal de las estimaciones individuales, siendo ésta

$$\hat{y} = \sum_{m=1}^{M} w_m f\left(x; \hat{a}_m\right), \; w_i \geq 0, i = 1, \ldots, M$$

que puede considerarse convexa si $\sum_{m=1}^{M} w_m = 1$.

Para una combinación de estimadores con restricciones, puede ser conveniente considerar como pesos

$$w_m = \frac{\frac{1}{n} \sum_{i \in S_m} L\left[y_i, c_{0m} + \sum_{l=1}^{L} c_{lm} f\left(x_i; \theta_{lm}\right) \right]}{\sum_{m=1}^{M} \frac{1}{n} \sum_{i \in S_m} L\left[y_i, c_{0m} + \sum_{l=1}^{L} c_{lm} f\left(x_i; \theta_{lm}\right) \right]}$$

siempre y cuando los estimadores individuales sean insesgados.

Es evidente que al perturbar habremos conseguido una disminución entre las covarianzas de los estimadores individuales y por tanto es de suponer que una reducción en el error cuadrático de estimación.

6.3.3 Combinación de estimadores sin restricciones

Una transformación en el espacio muestral (χ, B) provoca una perturbación en el espacio de medidas de probabilidad que denotaremos (Ω, A). Así pues, podemos tomar dos estimaciones de un punto en el espacio $y \in Y$ obtenidas mediante dos transformaciones en (χ, B), es decir

$$\hat{f}_1 = \hat{f}\left(x^1\right), \; \hat{f}_2 = \hat{f}\left(x^2\right) \text{ y } \; f \in \Im$$

Es evidente que se puede usar como función de perdida $L\left(Y, \hat{Y}\right) = Y - \hat{Y}$ y si llamamos error de predicción a $e^j = y - \hat{f}_j$, y además se verifica que

$$E\left[e^j\right] = 0, \; E\left[\left(e^j\right)^2\right] = \sigma^2 \text{ para } \; j = 1,2 \text{ y } \; E\left[e^1 e^2\right] = \ell \sigma^1 \sigma^2$$

En la situación así planteada, podemos considerar como estimación de y la combinación lineal de las dos predicciones

$$C = k\hat{f}_1 + (1-k)\hat{f}_2$$

Por tanto, el error de predicción de la combinación lineal tiene que ser

$$e^c = y - C = ke^1 + (1-k)e^2$$

y la varianza del error de predicción resulta ser

$$\sigma_c^2 = k^2\sigma_1^2 + (1-k)^2\sigma_2^2 + 2k(1-k)\ell\sigma_1\sigma_2$$

que es minimizada por

$$k^* = \frac{\sigma_2^2 - \ell\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\ell\sigma_1\sigma_2}$$

Así pues, resulta ser que

$$k^* \begin{matrix} > \\ = \\ < \end{matrix} \; sii \; \frac{\sigma_2}{\sigma_1} \begin{matrix} > \\ = \\ < \end{matrix} \; \ell$$

con lo cual, se verifica que $\sigma_{c,*}^2 < min(\sigma_1^2, \sigma_2^2)$ a menos que ℓ sea igual a σ_1/σ_2 o σ_2/σ_1.

En el caso general, si generalizamos a M transformaciones obtenemos M variables aleatorias $\{\hat{f}_j(x)\}_{j=1}^M$, que resultan ser M estimadores de un punto del espacio $y \in Y$. Notar que f_j pertenece a una familia fija $\Im$. Podríamos haber tomado el caso donde los estimadores son escogidos de $\Im_1, \ldots, \Im_j$ familias diferentes, esto es $f_j \in \Im_j$, $j \in \{1, \ldots, M\}$ y no realizar ninguna transformación en (χ, B).

Parece natural que la estimación final puede ser hecha utilizando una combinación lineal de las M predicciones $\hat{y} = \sum_{j=1}^M w_j \hat{f}_j(x)$, es evidente que el producto escalar es una aplicación continua y por tanto podemos utilizar una forma lineal continua sobre el espacio euclídeo, esto es, una combinación lineal de $\{\hat{f}_j(x)\}_{j=1}^M$.

En tal caso, la clase de combinaciones lineales, que puede ser un número infinito numerable de M estimadores obtenidos de una familia $\Im$, la denotaremos $lin_M(\Im)$, además se puede expresar como

$$Lin_M(\Im) = \left\{ f : f(x) = \sum_{j=1}^M w_j f_j(x), w_j \geq 0, f_j \in \Im \right\}$$

Por tanto el error cuadrático de estimación del estimador combinado puede descomponerse en

$$E\left[(\hat{y}-y)^2\right]=E\left[\left(w^t\hat{f}-E\left[w^t\hat{f}\right]\right)^2\right]+E\left(E\left[w^t\hat{f}\right]-y\right)^2$$
$$=w^t\Omega w+\left(w^t\mu-y\right)^2$$

siendo Ω la matriz de varianzas covarianzas, más concretamente $\Omega_{ij}=E\left[\left(\hat{f}_i-\mu_i\right)\left(\hat{f}_j-\mu_j\right)\right]$. Este resultado permite obtener dos partes; la primera corresponde a la varianza del funcional combinado, la segunda corresponde al sesgo del estimador combinado. Se verifica además que el mínimo riesgo esperado se obtiene para $w^*=\left(\mu\mu^t+\Omega\right)^{-1}y\mu$. Ahora bien, si tomamos $w_i=1/M$, se verifica

$$E\left[(\hat{y}-y)^2\right]=\frac{1}{M^2}\sum_{i=1}^{M}\Omega_{ii}+\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}\Omega_{ij}+\frac{1}{M^2}\left(\sum_{i=1}^{M}(\mu_i-y)\right)^2$$

Y además si denotamos por $\mu_i=media$, $\Omega_{ii}=\text{var}\,ianza$ y $\Omega_{ij}=\text{cov}$, permite obtener, más claramente, el siguiente resultado

$$E\left[(\hat{y}-y)^2\right]=\frac{1}{M}\,\text{var}\,ianza+\frac{M^2-M}{M^2}\,\text{cov}+(media-Y)^2$$

Como se puede observar, el sesgo del sistema combinado es idéntico al sesgo de cada estimador y la combinación no reduce este. En tal caso, habría que escoger los estimadores individuales con poco sesgo, conviene recordar que la regularización introduce sesgo. Por otro lado, la covarianza entre los estimadores debe ser pequeña, ya que esta no puede reducirse incrementando el número de estimadores en la combinación M. La varianza por su parte se ve reducida por un factor $1/M$.

Por tanto, como criterio debemos buscar estimadores individuales con poco sesgo y en consecuencia el estimador combinado tendrá poco sesgo, y con poca correlación entre ellos. En esta situación, el riesgo esperado del estimador combinado será significativamente menor que el de un estimador individual.

6.3.4 Combinación de estimadores con restricciones

Si añadimos al problema la restricción $\sum_{i=1}^{M}w_i=1$ siendo $w_i\geq 0\quad i=1,\ldots,M$, en este nuevo caso, la combinación es convexa. Así pues, la clase de combinaciones lineales convexa de orden M puede ser definida por el conjunto

$$Co_M(\Im)=\left\{f:f(x)=\sum_{j=1}^{M}w_j f_j(x),w_j\geq 0,\sum_{j=1}^{M}w_j=1,f_j\in\Im\right\}$$

Estamos ante un problema de optimización con restricciones, en este caso con una relación funcional. Como es sabido, éste puede ser transformado en una función lineal de la función objetivo y las restricciones, de esta manera el lagrangiano puede expresarse como

$$L = w^t \Omega w + \left(w^t \mu - y\right)^2 + \lambda\left(w^t u - 1\right)$$

si utilizamos una formulación alternativa para los pesos, esto es

$$w = \left(u^t g\right)^{-1} g \ , \ u = (1,\ldots,1)^t$$

Podemos resolver, igualando a cero las derivadas parciales, con lo que obtenemos $m+1$ ecuaciones para resolver $m+1$ variables. El óptimo se obtiene para

$$w^* = \left(\Omega + (\mu - yu)(\mu - yu)^t\right)^{-1} u$$

Finalmente, el error cuadrático medio de estimación puede expresarse como

$$E\left[(\hat{y} - y)^2\right] = \frac{1}{u'\left(\Omega + (\mu - yu)(\mu - yu)^t\right)^{-1} u}$$

Se verifica entonces que el estimador combinado es insesgado si los estimadores individuales son insesgados, puesto que los sesgos individuales aparecen explícitamente en el denominador. Además, si los estimadores individuales son incorrelados, el óptimo es

$$w_i^* = \frac{V[f_i]}{\sum_{i=1}^{M} V[f_i]}$$

En esta situación parece conveniente dar menos peso a aquellos estimadores que producen más error.

6.3.5 Convergencia del estimador combinado

La clase de combinaciones lineales $lin_M(\Im)$ o combinaciones lineales convexas $Co_M(\Im)$ de M estimadores bajo ciertas condiciones, como hemos visto, dan mejores resultados que estimadores individuales. Podemos, además, analizar la convergencia de su función de riesgo en función de la complejidad de los estimadores a través de la VC dimensión, del tamaño de la muestra de aprendizaje y del número de estimadores que intervienen en la combinación.

Si utilizamos la clase de combinaciones lineales, esto es $lin_M(\Im)$, de un número puede ser infinitamente numerable de estimadores de alguna familia fija $\Im$, suponiendo:

I. $z_n = \{(x_i, y_i)\}_{i=1}^n$ es una muestra aleatoria independiente de acuerdo a una distribución de probabilidad P.

II. $f_j : X \to \{+1, -1\}$, siendo $f = \Sigma_{j=1}^M w_j f_j \quad w_j \geq 0$.

III. $R(f) = P(yf(x) \leq 0)$.

En tal caso, existe una constante k tal que, con probabilidad al menos $1-\delta$, se puede establecer el siguiente resultado [25]

$$R(f) \leq \hat{R}(f) + k\sqrt{\frac{d\ln(n/d) + \ln(1/\delta)}{n}}$$

siendo $d = M\ln(M)VC\dim(\Im)$ y siempre que la VC-dimensión sea finita, $VC\dim(\Im) < \infty$. El segundo miembro depende de la VC dimensión de la clase de funcionales $\Im$ que, como hemos visto, es una medida de la complejidad de $\Im$. El segundo miembro también depende del tamaño de la muestra de aprendizaje que debería crecer al menos linealmente con el número de estimadores M introducidos en la combinación independientemente de $\hat{R}(f)$.

Si además, añadimos la restricción $\sum_{i=1}^m w_j = 1$ siendo $w_j \geq 0 \quad j = 1,\ldots,m$, estaremos ante la clase de combinaciones lineales convexas $Co_M(\Im)$ de un número que nuevamente, puede ser infinitamente numerable de estimadores de alguna familia fija $\Im$, que para un $\theta \in (0,1]$, se puede establecer el siguiente resultado [25]

$$R(f) \leq P_Z(yf(x) \leq \theta) + k\sqrt{\frac{1}{n}\frac{d\ln^2(n/d)}{\theta^2} + \ln(1/\delta)}$$

habida cuenta que $d = VC\dim(\Im)$. Acerca de este resultado cabe señalar que depende de la proporción del conjunto de aprendizaje con margen[6] menor que algún valor θ, siendo el margen $yf(x)$, de la VC dimensión de $\Im$, esta vez, penalizado por un parámetro θ y del tamaño de la muestra de aprendizaje. Notar que ahora no existe dependencia de M, el número de estimadores utilizados en la combinación.

Este resultado y el anterior usan la hipótesis de que la complejidad depende del estimador más complejo en la combinación, aun cuando éste puede tener un peso pequeño en esta. Está claro que un caso más realista sería si generalizamos al caso donde los clasificadores pueden ser escogidos de familias diferentes $\Im_1,\ldots,\Im_r$, las cuales pueden

tener VC dimensiones diferentes. En este caso, si $f_t \in \Im_{n_t}$ para $n_t \in \{1,\ldots,r\}$, si existe una constante k y para un $\theta \in (0,1]$ tal que con probabilidad al menos $1-\delta$, se puede establecer el siguiente resultado, LL.Mason, P. Bartlett, W. Sun Lee, M. Golea. (1998),

$$R(f) \le P_Z\left(yf(x) \le \theta\right) + k\sqrt{\frac{1}{n}\left(\frac{1}{\theta^2}\left(\mu \ln n + \ln r\right)\ln\left(n\theta^2/\mu\right) + \ln\left(1/\delta\right)\right)}^{1/2}$$

siendo $\mu = \sum_{t=1}^{M} w_t VC\dim\left(\Im_{n_t}\right)$ la media ponderada de la VC dimensión de los clasificadores.

Además, si tomamos un $\delta > 0$, existe una constante c tal que

$$R(f) \le 2P_Z\left(yf(x)\right) + \frac{1}{\theta^2}\sum_{i=1}^{n} w_i d_i B$$

siendo $B = \frac{c}{n} VC\dim(\Im)\log^2 n \log d$, diferentes elecciones de w_i y de θ darán diferentes estimaciones del riesgo de f.

6.3.6 Bagging

Si consideramos un estimador con funcional asociado T y F la distribución desconocida común de las variables que forman la muestra, podemos estimar $T(F)$ por $T(F_n)$, siendo F_n la bien conocida distribución empírica.

Además podemos realizar, como hemos visto, M transformaciones en el espacio muestral si utilizamos para ello M muestras Bootstrap independientes con la misma distribución que F_n, con lo que es posible generar $T_m(F_n)$ $m = 1,\ldots,M$ funcionales. Es sabido que si se cumple

$$T(F_n) \xrightarrow{\wp} T(F) \quad n \to \infty$$

el funcional T será consistente en F. De hecho, podemos pensar que cada uno de los M funcionales son consistentes, por tanto tiene sentido utilizar una sucesión consistente de estimadores y construir un estimador $T(F)$ por agregación/combinación, es decir que se puede construir un estimador de la forma $T^B(F_n) = \frac{1}{M}\sum_{m=1}^{M} T_m(F_n)$

---

[6] Si $y \in \{-1,1\}$, en este caso podemos coger como medida de la calidad de las estimaciones el margen, término en ingles "margin", esto es $yf(x)$. La función de perdida puede ser definida como $L(y,\hat{y}) = \max(0,1 - yf(x))$.

ya que resulta que $T^B(F_n) \overset{\wp}{\to} T(F)$.

De manera que Bagging[7] modifica la distribución muestral por reponderación aleatoria, por tanto

$$\hat{F}_m = \sum_{i=1}^{n} w_{im} \delta_{z_i} \ , \ m = \{1, \ldots, M\}$$

siendo los pesos $\{w_{im}\}_{i=1}^{n}$, al utilizar Bootstrap, obtenidos de una distribución multinomial y si $w_{im} \in \{0,1, \ldots, n\}$ con probabilidad $1/n$.

Si retomamos el modelo de predicción lineal visto anteriormente, los coeficientes utilizando Bagging son $\{\theta_{m0} = 1/M\}_{m=1}^{M}$. Por su parte el parámetro de regularización es de suponer que es $\lambda = \infty$, así pues la capacidad de aprendizaje del algoritmo no debería de incrementarse con el tamaño M de estimadores. Este resultado muestra que la sobreestimación debido a la complejidad del modelo no debería ser un problema en relación a cuando el número M de estimadores en la combinación aumenta.

Además es fácil ver que, si tenemos en cuenta los resultados obtenidos para estimadores combinados con restricciones, el estimador Bagging será insesgado si los estimadores individuales son insesgados. En esta situación y si suponemos que estos son a su vez incorrelados, podríamos sugerir que la elección óptima sería un estimador Bagging modificado, con expresión

$$T^{B'}(F_n) = \sum_{m=1}^{M} w_m T_m(F_n)$$

siendo

$$w_m = \frac{\text{var}[T_m(F_n)]}{\sum_{m=1}^{M} \text{var}[T_m(F_n)]}$$

En realidad, al disponer de una muestra finita, como ya se ha dicho, este resultado, en general, no es del todo cierto y solo podemos decir que se producirá una reducción de la correlación entre los estimadores individuales. Así pues se reducirá el error cuadrático de estimación del sistema combinado o de cualquier otra función de riesgo utilizada. Además la influencia de un dato anómalo será menor debido al uso de Boostrap.

[7] En ingles, Bootstrap sampling and aggregation.

6.3.7 Estimador Bagging para clasificación

Si y es un conjunto finito discreto, para un problema de clasificación la estimación se hará por simple voto, Breiman (1999). Por tanto

$$f_B(x) = \arg\max_j N_j,\ N_j = \#\{k;\ f(x, Z_k) = j\}$$

Naturalmente se verifica

I. La probabilidad de clasificación correcta de $f(x, Z)$ es

$$r = \sum_j \int P_Z(f(x, Z) = j) P(Y = j \mid X = x) P_X(dx).$$

II. La probabilidad de correcta clasificación de $f_A(x) = \arg\max_i P_Z(f(x, Z) = i)$ ha de ser

$$\sum_j \int I(\arg\max_i P_Z(f(x, Z) = i)) P(Y = j \mid X = x) P_X(dx).$$

III. Si el estimador óptimo es $f^*(x) = \arg\max_j P_Z(f(x, Z) = j)$, la probabilidad óptima de clasificación correcta es

$$\int \arg\max_j P_Z(f(x, Z) = j) P_X(dx).$$

IV. $C = \{x : f_A(x) = f^*(x)\}$.

En esta situación y si $P(C) \cong 1$, nos permite asegurar que $f_A(x)$ estará cercano al óptimo. Como era previsible, al igual que en regresión, con estimadores estables puede ser que el estimador Bagging no sea la elección óptima. Si en cambio lo será con estimadores inestables.

6.3.8 Random Forest

Random Forest, Breiman (2000), es un algoritmo que utiliza una combinación lineal convexa de árboles de decisión en los cuales se modifican por un lado la distribución conjunta de la muestra de aprendizaje y por otro el algoritmo utilizado para construir los árboles. En el primer caso, como hemos visto, se puede construir una familia de transformaciones G_t en (Ω, A) en el espacio paramétrico Θ. Una transformación en el espacio muestral (χ, B) provoca una perturbación en P_θ y por tanto en una parte del espacio de medidas de probabilidad en (Ω, A). Así pues, cada perturbación dará lugar a una estimación diferente de vectores de poda en cada una de las muestras.

En consecuencia, en RF las perturbaciones en P_θ se consiguen mediante muestras Bootstrap con reemplazamiento de la muestra de aprendizaje, con lo que se modifica la distribución muestral $F_n = 1/n \sum_{i=1}^{n} \delta_{z_i}$ por reponderación aleatoria, luego nuevamente $F_n^m = \sum_{i=1}^{n} w_{im} \delta_{z_i}$ para $m = 1, \ldots, M$ y los pesos $\{w_{im}\}_{i=1}^{n}$ siguen una distribución multinomial. Es evidente que cada F_n^m tiene la misma distribución que F_n y teóricamente cada muestra debe ser independiente una de otra.

En el segundo caso se produce una modificación del algoritmo que se usa en la construcción del árbol T. Desde luego, cada nodo define una partición del espacio euclídeo de la forma $x^q < a$, $x^q \geq a$ siendo $x \in X \subset \Re^q$ y cada nueva partición, notar que el árbol es orientado, genera dos nuevas particiones buscando sobre todo q y a la mejor estimación que por ejemplo en clasificación minimice el criterio de Gini o en regresión minimice la suma de cuadrados.

Cabe destacar que, en Random Forest, cada nueva partición no se realiza buscando sobre todo q sino sobre un subconjunto q_s elegido aleatoriamente de entre las q variables explicativas $1 \leq q_s \leq q$. Consecuentemente, cuando $q_s = q$, el algoritmo es equivalente a Bagging de árboles utilizando la metodología CART[8]. Por otro lado, si $q_s = 1$, se produciría una poda aleatoria. Sin embargo nada permite concluir que valor q_s es óptimo y si éste debe mantenerse durante todo el proceso de construcción de un árbol. En esta dirección Breiman sugiere tomar $q_s = \lfloor \log_2 q + 1 \rfloor$, notar que esta elección tiene como única consecuencia reducir los cálculos por un factor q_s / q.

Así pues, con estas dos modificaciones estamos introduciendo aleatoriedad en la construcción de cada uno de los árboles con lo que al combinar estos, deberá producirse una reducción en la estimación del riesgo esperado. La relación que existe entre el error del estimador combinado y el error de un estimador individual viene dada por Tumer, K. And Ghosh, J. (1999),

$$error_c = \frac{1 + \ell(M-1)}{M} error + error_{bayes}$$

[8] Existen diferentes algoritmos para construir árboles de decisión y las diferencias entre ellos están en la estrategia de podar los árboles, las reglas para particionar los nodos y el tratamiento de valores perdidos. Entre los principales algoritmos están: Árboles de Clasificación y Regresión (CART: Classification And Regression Tree). Otras metodologías son C4.5. Introducido por Quinlan, CHAID. "Chi-square automatic interaction detection", fue introducido por Kass, NewId. Es muy similar a C4.5, Arboles Bayesianos: Está basado en aplicación de métodos Bayesianos a arboles de decisión. Buntine, CN2. Introducido por Clark and Niblett (1988).

siendo ℓ el coeficiente de correlación entre los errores de los estimadores[9]. Si $\ell = 1$ el error del sistema combinado es igual al de un estimador simple. Si $\ell = 0$ el error del conjunto decrece proporcionalmente con el número de estimadores introducidos en la combinación, en nuestro caso árboles.

6.3.9 Convergencia de los árboles de decisión

Resultados de la teoría de la VC-dimensión sugieren que la muestra de aprendizaje debería crecer al menos linealmente con el tamaño del árbol $|T|$. De hecho, un resultado anterior muestra que el riesgo es inversamente proporcional al tamaño de la muestra y directamente proporcional a la complejidad de la familia de funcionales $\Im$ a través de la VC-dimensión. Conviene por tanto analizar que tipo de medida de la complejidad se puede utilizar para los árboles de decisión. En particular el tamaño del árbol no es siempre una buena medida de la complejidad y se da como alternativa una nueva medida. Ésta depende de la distribución inducida por la muestra de aprendizaje en los nodos terminales. Este resultado se sostiene bajo el resultado estudiado para una familia de combinaciones lineales convexas $Co_M(\Im)$.

Esta idea quedará reforzada con el siguiente resultado. Si consideramos un árbol de decisión definido como $T = \{t_k\}_{k=1,\ldots,K}$ siendo K el número de hojas del árbol, la estimación en el nodo terminal puede ser una combinación lineal convexa, con expresión

$$\sum_{k=1}^{k} w_k \gamma_k I(x \in R_k)$$

siendo $\gamma_k : X \to \{\overset{+}{-}1\}$ y $\sum_{k=1}^{K} w_k = 1$.

Por consiguiente, la estimación en el nodo terminal se puede realizar por voto mayoritario y por tanto la decisión en el nodo terminal puede ser del tipo $sign\left(\sum_{k=1}^{k} w_k \gamma_k I(x \in R_k)\right)$. Conviene en esta situación utilizar un resultado para la clase de combinaciones lineales convexas $Co_M(\Im)$, con lo que si tomamos una constante k y además un $\theta \in (0,1]$ con probabilidad al menos $1-\delta$, se cumple

$$R(f) \le \hat{R}(f) + \sum_{k=1}^{K} \hat{P}_k I(w_k \le \theta) + k\sqrt{\frac{1}{n}\left(\frac{1}{\theta^2}\left((v \ln n + \ln d)\ln\left(m\theta^2/v\right) + \ln(1/\delta)\right)\right)^{1/2}}$$

[9] la correlación entre $\hat{f}_i$ y $\hat{f}_j$ puede ser definida como la probabilidad de que $\hat{f}_i$ y $\hat{f}_j$ cometen el mismo error, es

Siempre, claro está, que la $VC\dim < \infty$, siendo $v = \bar{d}VC\dim(\Im)$ y d_k es el tamaño máximo del nodo terminal en el árbol T. No puede ser de otra forma, que $\bar{d}$ es la media ponderada del tamaño de los nodos terminales del árbol T, luego debe ser $\bar{d} = \Sigma_{k=1}^{K} w_k d_k$. Además, en cada nodo terminal $\{\hat{P}_k\}$ es la proporción de muestras de aprendizaje que son correctamente clasificadas por el nodo terminal k.

Luego el riesgo de un estimador basado en el voto mayoritario puede ser acotado en términos de la muestra de aprendizaje con un margen menor que un determinado θ más un término adicional. Este termino depende del número de muestras de aprendizaje, de alguna medida de la complejidad de $\Im$ y de θ, que en esta situación deberá ser elegido lo más grande posible.

6.3.10 Convergencia de una combinación de árboles

Como es lógico, podemos medir la complejidad de RF utilizando dos medidas, por un lado el tamaño de cada árbol, el cual debería de ser controlado para prevenir la sobreestimación. Y por otro, el número de árboles que se utilizan en la combinación. En el primer caso sería conveniente, para cada árbol en la combinación, seleccionar el árbol en T que minimiza por ejemplo $C(T) = \hat{L}_n(T) + \alpha|T|$. Sin embargo en RF no se utiliza ningún parámetro de regularización y los árboles crecen hasta su tamaño máximo.

Luego tanto el crecimiento de los árboles como el número de ellos en la combinación deberían de ser una fuente de sobreestimación y a pesar de todo el sobre ajuste no es un problema en RF. Consecuentemente, podemos pensar que se puede estar introduciendo cierta regularización debido a que cada nueva partición no se realiza buscando sobre todo q sino sobre un subconjunto $1 \le q_s \le q$ elegido aleatoriamente, así pues los árboles no estarían creciendo hasta su tamaño máximo.

Por tanto, puede observarse que si tomamos como en Bagging el modelo de predicción lineal, el parámetro de regularización tomaría un valor $\lambda < \infty$. No obstante, es de destacar que las estimaciones se

realizan utilizando Bagging y por tanto en este supuesto la capacidad de aprendizaje del algoritmo no debería de incrementarse con el número de árboles en la combinación y muy a nuestro pesar deberíamos considerar un valor $\lambda = \infty$.

Si nos centramos en el problema de la convergencia, un conjunto de árboles se combinan para formar una combinación lineal convexa, el riesgo depende en este caso de la proporción de muestras con pequeño margen y de la complejidad media de los árboles individuales.

De manera que, si existe una constante k tal que con probabilidad al menos $1-\delta$ y cada combinación F de árboles de decisión siendo $\theta_0 \in (0,1]$ y $\theta_t \in (0,1]$, se satisface

$$P(yF(x) \le 0) \le 2P(yF(x) \le \theta_0) + k\frac{1}{m}\left[\frac{\ln m}{\theta_0^2}\left(\sum_{t=1}^{T} w_t \min(C_1, C_2) + C_3\right) + \ln(1/\delta)\right]$$

$$C_1 = \frac{\ln(m/\theta_0)}{\theta_t^2}\left(\bar{d}_t VC\dim(\mu)\ln m + \ln\left(T_t \max_l d_l\right)\right) + \frac{m\theta_0}{\ln m}\sum_{s=1}^{T_t} P_{t,s}\,1(v_{ts} < \theta_t)$$

$$C_2 = T_t\left(VC\dim(\mu)\ln m + \ln\left(\max_l d_l\right)\right)$$

$$C_3 = \ln\left(\frac{T\ln(m/\theta_0)}{\min_t \theta_t^2}\right)$$

donde μ representa la clase de funciones utilizadas en los nodos terminales. Deberemos tomar un valor de θ_0 lo mas grande posible para minimizar el segundo termino.

6.3.11 El margen

Una vez más, si suponemos que $Y \in \{-1,1\}$ y tomamos un estimador de la forma $sign(f(x))$, el margen puede ser definido como $yf(x)$. Además si tenemos $f_1(x), f_2(x), \ldots, f_M(x)$, un camino para combinar los estimadores individuales es utilizando la clase de combinaciones lineales de un número puede ser que infinitamente numerable de estimadores de alguna familia fija $\Im$.

Así pues, podemos definir un estimador, el cual podemos llamar de voto mayoritario, de la forma $sign\sum_{m=1}^{M} w_m f_m(x)$, siendo $f_m : X \to \{-1,1\}$ y como $w_m = 1/M$ $\forall m = 1, \ldots, M$ estaremos ante la clase de combinaciones lineales convexas $Co_M(\Im)$.

Desde luego, se puede notar que en RF cada árbol es escogido de diferentes familias $\Im_1,\ldots,\Im_M$ que pueden tener complejidades diferentes. En definitiva cada árbol no se selecciona de una familia fija sino de familias que, dependiendo como midamos la complejidad, pueden tener VC-dimensiones diferentes, siendo pues $f_m \in \Im_{n_m}$ para $n_m \in \{1,\ldots,M\}$.

Luego si, $f_1(x), f_2(x),\ldots, f_M(x)$ es un conjunto de clasificadores entonces una definición alternativa del margen es

$$mg(X,Y)=\frac{1}{M}\sum_{m=1}^{M} I(f_m(X)=Y)-\max_{j\neq y}\frac{1}{M}\sum_{m=1}^{M} I(f_m(X)=j)$$

entonces, es evidente que:

- Si $mg(X,Y)>0$ el conjunto de clasificadores vota por la clase correcta.
- Si $mg(X,Y)<0$ el conjunto de clasificadores vota por la clase incorrecta.

Por tanto, el riesgo del estimador debe ser

$$R(f)=P_{X,Y}(mg(X,Y)<0)$$

Además, como cada árbol depende de un vector de parámetros θ, de forma que el estimador puede ser definido como $f_m(X,\theta_m)$ y en esta nueva situación el margen tomará la forma

$$mg(X,Y)=\frac{1}{M}\sum_{m=1}^{M} I(f_m(X,\theta_m)=Y)-\max_{j\neq y}\frac{1}{M}\sum_{m=1}^{M} I(f_m(X,\theta_m)=j)$$

Claro está, que la sucesión de estimadores $1/M\sum_{m=1}^{M} I(f_m(X,\theta)=Y)$ asociada a diferentes tamaños de M será consistente para estimar una función del parámetro $P_\theta(y=f(x,\theta))$ si cuando $M\to\infty$, se verifica

$$\frac{1}{M}\sum_{m=1}^{M} I\left(y=f\left(x,\theta_m\right)\right)\xrightarrow{\wp_\theta} P_\theta(y=f(x,\theta))$$

Respectivamente, se puede sustituir la convergencia en probabilidad por la convergencia casi segura, así pues se puede mostrar, Breiman (1999), que existe un conjunto de probabilidad nula en la sucesión de $\theta_1,\theta_2,\ldots$, tal que fuera de este para todo x

$$1/N\sum_{i=1}^{N} I\left(y=f\left(x,\theta_i\right)\right)\xrightarrow{\wp_\theta-c.s.} P_\theta(y=f(x,\theta))$$

siendo $f(x,\theta)$ la unión de un conjunto de hiperrectangulos R_k, si tomamos una muestra, el número de estimadores será finito $K<\infty$ y definimos N_k como el número de veces que $\varphi(\theta)=k$ si $\{x/\ f(x,\theta)=y\}=R_k$, se verifica

$$\frac{1}{N}\sum_{n=1}^{N} I(f(\theta_n,x)=y)=\frac{1}{N}\sum_{k} N_k I(x\in S_k)$$

por la ley fuerte de los grandes números

$$N_k=\frac{1}{N}\sum_{i=1}^{N} I(\varphi(\theta_n)=k)\overset{P\theta-c.s.}{\rightarrow} P_\theta(\varphi(\theta_n)=k)$$

tomando la unión de todos los conjuntos en los cuales la convergencia no ocurre para algún valor de k dado un conjunto de probabilidad nula, la unión de una cantidad numerable de conjuntos de medida nula es un conjunto de medida nula y la probabilidad fuera de el es

$$\frac{1}{N}\sum_{i=1}^{N} I(f(x,\theta_n)=y)\overset{P\theta-c.s.}{\rightarrow}\sum_k P_\theta(\varphi(\theta_n)=k)I(x\in R_k)=P_\theta(f(x,\theta)=y)$$

En consecuencia, podemos admitir la convergencia casi segura del riesgo muestral al riesgo teórico, con lo que para toda sucesión $\theta_1,\theta_2,\ldots$ el riesgo muestral converge a

$$P_{X,Y}(mg(X,Y)<0)\overset{\wp_\theta-c.s.}{\rightarrow} P_{X,Y}\left(P_\theta(f(x,\theta)=Y)-\max_{j\neq Y}P_\theta(f(x,\theta)=j)<0\right)$$

y por tanto, la interpretación de este resultado es clara, la sobreestimación no debería ser un problema en RF aunque se sigan introduciendo estimadores o sea árboles, en la combinación.

6.3.12 Riesgo del estimador del margen

El voto mayoritario es $sign\sum_{m=1}^{M} w_m f_m(x)$ siendo $f_m: X\rightarrow\{-1,1\}$ y además como $w_m=1/M$, $\forall m=1,\ldots,M$ estaremos ante la clase de combinaciones lineales de estimadores que pueden tener VC-dimensiones diferentes $f_m\in\Im_{n_m}$ para $n_m\in\{1,\ldots,M\}$. En esta situación, la cota para el riesgo teórico se incrementará si se incrementa la complejidad media del bosque, siendo $\sum_{t=1}^{M} w_t VC\dim(\Im_{n_t})$ la media ponderada de la VC-dimensión de los estimadores individuales.

Notar que, en un árbol, el aumento de la complejidad puede reducir sesgo pero sin embargo puede aumentar la varianza por otro lado depende de la proporción de muestras de

aprendizaje con margen menor que un determinado θ. Además, el riesgo se reducirá si se incrementa la muestra de aprendizaje.

Por tanto, mediante los resultados vistos hasta ahora, la complejidad media del bosque no se incrementa ya que parece ser, bajo hipótesis, que RF utiliza cierta regularización para controlar el tamaño de los árboles. Lo cual, muy a nuestro pesar, puede ser a costa de introducir cierto sesgo en su construcción.

Por su parte el número de árboles introducido en la combinación no produce sobreestimación al estar asegurada la convergencia casi segura del riesgo muestral.

Al igual como hemos visto para combinación de estimadores, el estimador combinado será insesgado si los estimadores individuales son insesgados y el riesgo muestral se reducirá si los árboles son incorrelados entre si, luego si

$$mr(X,Y) = P_\theta(f(X,\theta) = Y) - \max_{j \neq Y} P_\theta(f(X,\theta) = j)$$

la estrategia de RF puede ser la de buscar árboles que den la clase correcta en media, esto es árboles que tengan un valor de $s = E_{X,Y}[mr(X,Y)]$ alto, entonces si utilizamos la desigualdad de Chebychev's

$$P(|mg(X,Y) - E[mg(X,Y)]| > E[mg(X,Y)]) < \frac{V[mg(X,Y)]}{E[mg[X,Y]]^2}$$

con lo cual

$$\hat{R}(f) \leq \frac{V[mr(X,Y)]}{s^2}$$

Si tomamos $\hat{j}(X,Y) = \max_{j \neq Y} P\theta(h(X,\theta) = j)$, donde $\hat{j}(X,Y)$ es la clase predicha más probable distinta de *Y*, según ello podemos definir

$$rmg(\theta, X, Y) = I(h(X,\theta) = Y) - \max_{j \neq Y} I(h(X,\theta) = \hat{j}(X,Y))$$

notar que $mg(X,Y) = E_\theta[rmg(X,Y)]$, se verifica

$$\begin{aligned} \text{var}(mr) &= E_{\theta,\theta'}[\text{cov}_{X,Y}[rmg(\theta, X, Y) rmg(\theta', X, Y)]] \\ &= E_{\theta,\theta'}[\ell(\theta,\theta') sd(\theta) sd(\theta')] \end{aligned}$$

si $\ell(\theta,\theta')$ es la correlación entre $rmg(\theta, X, Y)$ y $rmg(\theta', X, Y)$, entonces

$$\begin{aligned} \text{var}(mr) &= \bar{\ell}(E_\theta[sd(\theta)])^2 \\ &\leq \bar{\ell}(E_\theta[\text{var}(\theta)]) \end{aligned}$$

$$E_\theta[\text{var}(\theta)] \le E_\theta\left[E_{X,Y}[rmg(\theta, X, Y)]^2 - s^2\right]$$
$$\le 1 - s^2$$

luego [2]

$$\hat{R}(f) \le \frac{\bar{\ell}\left(1 - s^2\right)}{s^2}$$

siendo $\bar{\ell}$ la correlación media ponderada sobre todos los posibles pares (θ, θ').

Este resultado muestra que la correlación entre estimadores debe ser pequeña, esto se debe ya el estimador RF es una combinación lineal de estimadores, la correlación no se reducirá conforme aumente el número de árboles M en el bosque. De hecho, RF introduce aleatoriedad en la construcción de los árboles en el bosque, con lo que se reducirá la correlación entre ellos. La varianza por su parte disminuirá conforme aumente el número de estimadores en la combinación.

7. Conclusiones

Como hemos visto los test de bondad de ajuste para series temporales se basan en la secuencia de autocorrelaciones o bien dependen funcionalmente de la función de autocovarianzas o de la función autocorrelaciones. Lo mismo ocurre con los test de raíces unitarias como es el caso del test de Dickey-Fuller de raíces unitarias o el test KPSS de estacionariedad.

Bajo este contexto, el coeficiente de correlación de Pearson no es una medida robusta entre dos variables, la estimación puede verse afectada por un outlier. En esta situación y como se ha mostrado el test de Dickey-Fuller no es robusto en presencia de un outlier aditivo. Como consecuencia directa tenemos que la distribución asintótica del estadístico de Dickey-Fuller es sesgada a la derecha y consecuentemente se sobre rechazara la hipótesis de raíz unitaria a favor de la alternativa estacionaria

Una alternativa son los métodos robustos de estimación que han mostrado, tal y como muestran los resultados presentados en las tablas del capitulo 5, ser preferibles a la estimación clásica, aún a pesar del sesgo que presentan en favor de la hipótesis de estacionariedad.

Bajo condiciones sin presencia de OA, tablas 5.1, 5.2 y 5.3, se ha mostrado, empíricamente, que todos los estimadores basados en métodos robustos son más potentes que el estimador clásico para tamaños de muestra de 50, 100 y 200, y casi para todos los niveles de significación. Sin embargo estos rechazarán la hipótesis de raíz unitaria en un mayor número de ocasiones que el estimador clásico.

En presencia de un OA, tablas 5.1, 5.2 y 5.3, se ha mostrado, que los estimadores robustos practicamente no se ven afectados en su potencia, si en cambio lo será el estimador clásico que rechazará la presencia de una raíz unitaria en prácticamente casi el 100% de las simulaciones, para cualquier tamaño de muestra y cualquier nivel de significación.

El estimador basado en una combinación de los estimadores robustos individuales parece ser no tan potente como estos, sin embargo, cabe decir, en su favor, que rechazará la presencia de una raíz unitaria en un porcentaje menor que estos. Supone pues, una mejora respecto a estos utilizados de forma individual y en presencia de un outlier aditivo es preferible, sin lugar a dudas, al estadístico de Dickey-Fuller, tal y como confirman los resultados obtenidos y presentados en las tablas 6.1, 6.2, 6.3, 6.4 y 6.5.

En esta situación dada la idoneidad, demostrada empíricamente, del uso de estimadores robustos y en particular de una combinación de estos bajo algún tipo de algoritmo basado en aprendizaje estadístico, en esta situación, sería útil el estudio de estos métodos en test como pueden ser el test KPSS o en el problema del estudio de test de raíces unitarias en procesos con tendencia.

8. Anexos

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,331	0,537	0,522	0,506	0,571	0,524
0,85	0,177	0,383	0,355	0,416	0,394	0,386
0,9	0,072	0,260	0,226	0,338	0,256	0,273
0,95	0,026	0,157	0,121	0,261	0,148	0,182
1	0,005	0,096	0,066	0,204	0,081	0,128
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,936	0,947	0,949	0,885	0,967	0,914
0,85	0,697	0,819	0,812	0,745	0,852	0,777
0,9	0,317	0,536	0,500	0,541	0,559	0,518
0,95	0,069	0,260	0,221	0,325	0,251	0,281
1	0,009	0,098	0,076	0,180	0,088	0,139
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	0,999	1,000	1,000
0,85	1,000	0,999	0,999	0,995	0,999	0,996
0,9	0,926	0,947	0,945	0,914	0,963	0,921
0,95	0,319	0,540	0,514	0,558	0,573	0,524
1	0,010	0,089	0,068	0,161	0,079	0,135

Tabla 8.1: Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.01$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,989	0,520	0,671	0,493	0,538	0,477
0,85	0,988	0,400	0,573	0,414	0,394	0,379
0,9	0,987	0,262	0,440	0,326	0,251	0,273
0,95	0,991	0,159	0,309	0,245	0,147	0,164
1	0,991	0,104	0,228	0,194	0,084	0,131
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,993	0,949	0,971	0,881	0,969	0,917
0,85	0,995	0,823	0,896	0,764	0,849	0,786
0,9	0,996	0,559	0,705	0,554	0,580	0,535
0,95	0,994	0,257	0,444	0,317	0,256	0,266
1	0,993	0,091	0,225	0,158	0,088	0,115
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,998	1,000	1,000	1,000	1,000	1,000
0,85	0,996	0,997	0,999	0,994	1,000	0,995
0,9	0,998	0,940	0,973	0,912	0,958	0,924
0,95	0,995	0,550	0,714	0,550	0,567	0,524
1	0,992	0,100	0,229	0,179	0,096	0,145

Tabla 8.2: Con outlier aditivo. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.01$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,911	0,934	0,932	0,864	0,954	0,901
0,85	0,147	0,354	0,325	0,396	0,367	0,348
0,9	0,059	0,262	0,212	0,324	0,246	0,258
0,95	0,021	0,177	0,135	0,268	0,152	0,188
1	0,004	0,101	0,078	0,211	0,087	0,128
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,911	0,934	0,932	0,864	0,954	0,901
0,85	0,664	0,795	0,784	0,712	0,822	0,753
0,9	0,294	0,528	0,493	0,508	0,541	0,491
0,95	0,080	0,269	0,226	0,342	0,264	0,281
1	0,007	0,098	0,071	0,192	0,088	0,147
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000
0,85	0,999	0,997	0,996	0,991	0,998	0,993
0,9	0,890	0,928	0,923	0,895	0,945	0,900
0,95	0,297	0,524	0,494	0,538	0,552	0,513
1	0,009	0,095	0,073	0,169	0,086	0,142

Tabla 8.3: Perturbando el espacio muestral. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.01$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,995	0,944	0,972	0,878	0,956	0,911
0,85	0,964	0,386	0,538	0,403	0,382	0,364
0,9	0,983	0,523	0,682	0,515	0,538	0,492
0,95	0,966	0,159	0,299	0,244	0,142	0,169
1	0,970	0,093	0,232	0,168	0,083	0,105
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,995	0,944	0,972	0,878	0,956	0,911
0,85	0,992	0,795	0,882	0,727	0,826	0,752
0,9	0,983	0,523	0,682	0,515	0,538	0,492
0,95	0,986	0,257	0,452	0,317	0,259	0,262
1	0,979	0,100	0,224	0,186	0,087	0,137
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,998	1,000	0,998	1,000	1,000	1,000
0,85	0,997	0,997	0,996	0,992	0,998	0,994
0,9	0,998	0,948	0,972	0,901	0,962	0,918
0,95	0,996	0,530	0,693	0,538	0,545	0,517
1	0,989	0,099	0,240	0,170	0,090	0,140

Tabla 8.4: Con outlier aditivo y perturbando el espacio muestral. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.01$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,580	0,748	0,721	0,644	0,768	0,705
0,85	0,363	0,595	0,551	0,573	0,614	0,558
0,9	0,173	0,459	0,407	0,464	0,456	0,439
0,95	0,065	0,302	0,259	0,364	0,290	0,303
1	0,016	0,199	0,144	0,299	0,167	0,220
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,989	0,988	0,990	0,951	0,992	0,974
0,85	0,888	0,935	0,925	0,880	0,946	0,903
0,9	0,563	0,756	0,711	0,705	0,752	0,711
0,95	0,167	0,449	0,387	0,477	0,438	0,437
1	0,023	0,196	0,153	0,292	0,174	0,224
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	0,999	0,998	1,000	1,000
0,9	0,987	0,988	0,990	0,974	0,994	0,975
0,95	0,565	0,758	0,713	0,728	0,767	0,720
1	0,021	0,194	0,147	0,274	0,174	0,229

Tabla 8.5: Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.025$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,989	0,739	0,818	0,651	0,755	0,692
0,85	0,988	0,611	0,746	0,550	0,618	0,578
0,9	0,987	0,461	0,618	0,462	0,450	0,435
0,95	0,991	0,308	0,480	0,375	0,286	0,300
1	0,991	0,204	0,380	0,285	0,169	0,221
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,993	0,991	0,990	0,948	0,995	0,984
0,85	0,995	0,934	0,957	0,888	0,942	0,910
0,9	0,996	0,767	0,844	0,715	0,777	0,720
0,95	0,994	0,440	0,622	0,465	0,441	0,436
1	0,993	0,186	0,368	0,258	0,173	0,206
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,998	1,000	1,000	1,000	1,000	1,000
0,85	0,996	1,000	0,999	1,000	1,000	0,999
0,9	0,998	0,985	0,991	0,968	0,991	0,978
0,95	0,996	0,753	0,846	0,724	0,762	0,713
1	0,992	0,202	0,385	0,285	0,185	0,245

Tabla 8.6: Con outlier aditivo. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.025$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,575	0,736	0,718	0,636	0,758	0,687
0,85	0,314	0,564	0,523	0,542	0,571	0,544
0,9	0,161	0,437	0,395	0,455	0,432	0,414
0,95	0,060	0,326	0,291	0,380	0,321	0,335
1	0,012	0,210	0,167	0,306	0,186	0,221
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,987	0,979	0,981	0,941	0,987	0,969
0,85	0,873	0,919	0,910	0,847	0,935	0,880
0,9	0,537	0,722	0,696	0,685	0,746	0,690
0,95	0,171	0,460	0,400	0,495	0,449	0,430
1	0,015	0,204	0,154	0,300	0,187	0,242
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	0,999	0,998	0,999	0,998
0,9	0,988	0,981	0,977	0,967	0,986	0,968
0,95	0,541	0,740	0,700	0,721	0,756	0,709
1	0,023	0,199	0,151	0,280	0,170	0,238

Tabla 8.7: Perturbando el espacio muestral. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.025$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,974	0,723	0,801	0,634	0,750	0,655
0,85	0,968	0,577	0,702	0,546	0,589	0,549
0,9	0,976	0,444	0,614	0,460	0,435	0,435
0,95	0,967	0,302	0,472	0,356	0,282	0,292
1	0,970	0,199	0,381	0,267	0,167	0,199
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,995	0,979	0,988	0,946	0,989	0,966
0,85	0,993	0,928	0,951	0,867	0,940	0,891
0,9	0,986	0,709	0,818	0,682	0,720	0,686
0,95	0,987	0,452	0,624	0,479	0,436	0,428
1	0,980	0,198	0,374	0,294	0,173	0,233
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,998	1,000	0,998	1,000	1,000	1,000
0,85	0,997	0,999	0,997	0,998	1,000	0,999
0,9	0,999	0,985	0,990	0,971	0,991	0,978
0,95	0,998	0,726	0,852	0,711	0,741	0,696
1	0,989	0,202	0,383	0,286	0,179	0,253

Tabla 8.8: Con outlier aditivo y perturbando el espacio muestral. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.025$.

n=50			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,782	0,868	0,861	0,762	0,893	0,837
0,85	0,578	0,770	0,746	0,687	0,791	0,726
0,9	0,317	0,629	0,577	0,571	0,620	0,593
0,95	0,124	0,462	0,400	0,479	0,440	0,438
1	0,034	0,325	0,255	0,409	0,289	0,323
n=100			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,997	0,997	0,996	0,984	0,998	0,993
0,85	0,967	0,974	0,969	0,944	0,977	0,959
0,9	0,762	0,876	0,843	0,818	0,879	0,837
0,95	0,293	0,626	0,560	0,620	0,609	0,595
1	0,043	0,325	0,248	0,395	0,285	0,339
n=200			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	1,000	1,000	1,000	1,000
0,9	1,000	0,998	0,997	0,986	0,998	0,995
0,95	0,771	0,882	0,858	0,850	0,887	0,851
1	0,041	0,311	0,243	0,395	0,282	0,340

Tabla 8.9: Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.05$.

n=50			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,990	0,867	0,896	0,773	0,884	0,842
0,85	0,988	0,764	0,839	0,678	0,767	0,729
0,9	0,988	0,630	0,743	0,593	0,616	0,593
0,95	0,991	0,467	0,625	0,488	0,445	0,449
1	0,991	0,313	0,508	0,383	0,292	0,338
n=100			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,993	0,999	0,991	0,985	1,000	0,994
0,85	0,995	0,977	0,978	0,954	0,981	0,961
0,9	0,996	0,893	0,917	0,839	0,902	0,862
0,95	0,994	0,625	0,752	0,614	0,610	0,587
1	0,995	0,301	0,506	0,368	0,273	0,311
n=200			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,998	1,000	1,000	1,000	1,000	1,000
0,85	0,996	1,000	0,999	1,000	1,000	1,000
0,9	0,998	0,996	0,994	0,986	0,997	0,992
0,95	0,996	0,880	0,920	0,847	0,886	0,845
1	0,993	0,323	0,524	0,391	0,300	0,359

Tabla 8.10: Con outlier aditivo. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.05$.

n=50			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,990	0,863	0,892	0,773	0,884	0,842
0,85	0,514	0,741	0,691	0,663	0,745	0,696
0,9	0,296	0,611	0,557	0,565	0,606	0,572
0,95	0,124	0,488	0,422	0,498	0,461	0,461
1	0,033	0,342	0,283	0,399	0,319	0,338
n=100			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,998	0,993	0,993	0,972	0,997	0,988
0,85	0,962	0,974	0,970	0,929	0,981	0,957
0,9	0,757	0,852	0,842	0,809	0,872	0,817
0,95	0,294	0,632	0,565	0,622	0,616	0,605
1	0,036	0,340	0,271	0,418	0,301	0,359
n=200			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	1,000	1,000	1,000	1,000
0,9	0,999	0,995	0,994	0,988	0,998	0,988
0,95	0,737	0,868	0,831	0,839	0,868	0,830
1	0,045	0,313	0,247	0,389	0,278	0,337

Tabla 8.11: Perturbando el espacio muestral. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.05$.

n=50			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,990	0,853	0,892	0,775	0,880	0,833
0,85	0,973	0,736	0,806	0,663	0,742	0,701
0,9	0,984	0,650	0,764	0,543	0,646	0,533
0,95	0,967	0,450	0,607	0,469	0,417	0,428
1	0,971	0,312	0,512	0,371	0,288	0,307
n=100			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,996	0,995	0,992	0,979	0,997	0,989
0,85	0,994	0,975	0,974	0,929	0,978	0,960
0,9	0,987	0,854	0,900	0,793	0,870	0,811
0,95	0,989	0,615	0,747	0,613	0,604	0,580
1	0,981	0,322	0,526	0,402	0,294	0,351
n=200			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,998	1,000	0,998	1,000	1,000	1,000
0,85	0,998	1,000	0,998	1,000	1,000	1,000
0,9	0,999	0,996	0,993	0,992	0,997	0,991
0,95	0,999	0,868	0,924	0,831	0,873	0,828
1	0,989	0,338	0,531	0,416	0,307	0,374

Tabla 8.12: Con outlier aditivo y perturbando el espacio muestral. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.05$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,880	0,921	0,921	0,819	0,944	0,900
0,85	0,710	0,853	0,828	0,770	0,857	0,812
0,9	0,442	0,715	0,677	0,644	0,716	0,686
0,95	0,185	0,575	0,508	0,553	0,543	0,541
1	0,051	0,429	0,343	0,485	0,380	0,421
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,999	0,999	0,999	0,991	1,000	0,996
0,85	0,986	0,986	0,982	0,969	0,989	0,977
0,9	0,868	0,929	0,909	0,876	0,940	0,899
0,95	0,417	0,730	0,664	0,697	0,717	0,682
1	0,065	0,424	0,338	0,475	0,384	0,428
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	1,000	1,000	1,000	1,000
0,9	1,000	0,999	0,999	0,993	0,999	0,999
0,95	0,877	0,930	0,913	0,908	0,936	0,905
1	0,061	0,416	0,341	0,478	0,382	0,433

Tabla 8.13: Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.075$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,990	0,919	0,929	0,829	0,931	0,891
0,85	0,988	0,830	0,885	0,759	0,843	0,808
0,9	0,988	0,721	0,819	0,669	0,721	0,686
0,95	0,991	0,575	0,706	0,568	0,549	0,543
1	0,991	0,407	0,605	0,447	0,381	0,408
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,993	1,000	0,992	0,995	1,000	0,998
0,85	0,995	0,992	0,984	0,972	0,993	0,980
0,9	0,996	0,942	0,943	0,903	0,942	0,912
0,95	0,994	0,736	0,817	0,699	0,720	0,687
1	0,995	0,401	0,594	0,452	0,361	0,388
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,998	1,000	1,000	1,000	1,000	1,000
0,85	0,996	1,000	0,999	1,000	1,000	1,000
0,9	0,998	0,998	0,995	0,994	0,999	0,996
0,95	0,996	0,928	0,942	0,898	0,930	0,906
1	0,994	0,433	0,607	0,478	0,394	0,443

Tabla 8.14: Con outlier aditivo. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.075$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,845	0,902	0,915	0,802	0,932	0,879
0,85	0,662	0,835	0,797	0,738	0,830	0,791
0,9	0,415	0,702	0,661	0,644	0,705	0,667
0,95	0,181	0,579	0,515	0,571	0,561	0,552
1	0,053	0,436	0,375	0,473	0,405	0,425
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	0,997	0,999	0,987	0,999	0,993
0,85	0,991	0,986	0,987	0,955	0,992	0,978
0,9	0,856	0,913	0,899	0,864	0,923	0,877
0,95	0,413	0,735	0,677	0,709	0,723	0,705
1	0,053	0,426	0,344	0,501	0,386	0,439
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	1,000	1,000	1,000	1,000
0,9	1,000	0,998	0,999	0,993	1,000	0,993
0,95	0,849	0,920	0,893	0,892	0,924	0,895
1	0,070	0,401	0,327	0,463	0,364	0,418

Tabla 8.15: Perturbando el espacio muestral. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.075$.

n=50 Φ	Clásico	Higly(Coman)	Population Percentage	Biweight Midcovariance	MCD	MVE
0,8	0,980	0,910	0,912	0,819	0,923	0,841
0,85	0,977	0,821	0,854	0,742	0,830	0,780
0,9	0,974	0,712	0,803	0,654	0,712	0,680
0,95	0,970	0,565	0,694	0,541	0,536	0,531
1	0,971	0,405	0,597	0,441	0,380	0,386
n=100 Φ	Clásico	Higly(Coman)	Population Percentage	Biweight Midcovariance	MCD	MVE
0,8	0,996	0,997	0,993	0,989	0,998	0,995
0,85	0,994	0,987	0,980	0,960	0,988	0,977
0,9	0,989	0,917	0,936	0,856	0,923	0,884
0,95	0,990	0,728	0,818	0,694	0,714	0,680
1	0,981	0,428	0,614	0,477	0,381	0,426
n=200 Φ	Clásico	Higly(Coman)	Population Percentage	Biweight Midcovariance	MCD	MVE
0,8	0,998	1,000	0,998	1,000	1,000	1,000
0,85	0,998	1,000	0,998	1,000	1,000	1,000
0,9	0,999	0,999	0,995	0,997	1,000	0,996
0,95	0,999	0,923	0,947	0,886	0,924	0,897
1	0,989	0,427	0,602	0,495	0,388	0,457

Tabla 8.16: Con outlier aditivo y perturbando el espacio muestral. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.075$.

n=50 Φ	Clásico	Higly(Coman)	Population Percentage	Biweight Midcovariance	MCD	MVE
0,8	0,930	0,951	0,950	0,863	0,966	0,934
0,85	0,790	0,897	0,875	0,815	0,910	0,860
0,9	0,537	0,789	0,753	0,700	0,792	0,748
0,95	0,243	0,652	0,593	0,612	0,630	0,622
1	0,076	0,499	0,418	0,545	0,459	0,492
n=100 Φ	Clásico	Higly(Coman)	Population Percentage	Biweight Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	0,995	1,000	0,998
0,85	0,998	0,990	0,990	0,980	0,994	0,984
0,9	0,932	0,953	0,945	0,914	0,960	0,933
0,95	0,519	0,801	0,752	0,754	0,793	0,757
1	0,090	0,499	0,416	0,544	0,455	0,492
n=200 Φ	Clásico	Higly(Coman)	Population Percentage	Biweight Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	1,000	1,000	1,000	1,000
0,9	1,000	0,999	0,999	0,998	1,000	1,000
0,95	0,934	0,958	0,942	0,935	0,958	0,936
1	0,085	0,503	0,419	0,537	0,458	0,499

Tabla 8.17: Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.1$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,990	0,950	0,946	0,865	0,956	0,926
0,85	0,988	0,884	0,908	0,807	0,895	0,858
0,9	0,988	0,790	0,861	0,720	0,786	0,758
0,95	0,991	0,652	0,762	0,620	0,626	0,618
1	0,991	0,485	0,664	0,502	0,459	0,476
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,993	1,000	0,993	0,997	1,000	0,999
0,85	0,995	0,995	0,985	0,982	0,997	0,989
0,9	0,996	0,963	0,955	0,928	0,964	0,945
0,95	0,994	0,803	0,853	0,759	0,788	0,762
1	0,995	0,481	0,654	0,512	0,437	0,465
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,998	1,000	1,000	1,000	1,000	1,000
0,85	0,996	1,000	0,999	1,000	1,000	1,000
0,9	0,998	0,998	0,996	0,996	1,000	0,997
0,95	0,996	0,950	0,956	0,930	0,950	0,933
1	0,994	0,512	0,669	0,547	0,471	0,521

Tabla 8.18: Con outlier aditivo. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.1$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,912	0,942	0,943	0,854	0,954	0,923
0,85	0,762	0,884	0,848	0,788	0,879	0,845
0,9	0,517	0,766	0,738	0,693	0,770	0,732
0,95	0,239	0,654	0,597	0,635	0,637	0,619
1	0,076	0,507	0,442	0,528	0,468	0,488
n=100			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	0,998	0,999	0,991	0,999	0,995
0,85	0,996	0,992	0,994	0,975	0,995	0,987
0,9	0,922	0,937	0,931	0,902	0,945	0,911
0,95	0,514	0,801	0,747	0,772	0,779	0,772
1	0,076	0,513	0,420	0,557	0,461	0,501
n=200			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	1,000	1,000	1,000	1,000	1,000	1,000
0,85	1,000	1,000	1,000	1,000	1,000	1,000
0,9	1,000	1,000	1,000	0,995	1,000	0,997
0,95	0,909	0,944	0,929	0,924	0,947	0,922
1	0,088	0,472	0,399	0,522	0,432	0,486

Tabla 8.19: Perturbando el espacio muestral. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.1$.

n=50			Population	Biweight		
Φ	Clásico	Higly(Coman)	Percentage	Midcovariance	MCD	MVE
0,8	0,980	0,956	0,956	0,857	0,955	0,935
0,85	0,982	0,870	0,884	0,784	0,875	0,833
0,9	0,978	0,787	0,846	0,712	0,776	0,738
0,95	0,972	0,638	0,752	0,604	0,614	0,594
1	0,971	0,486	0,653	0,510	0,456	0,458
n=100			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovarianc	MCD	MVE
0,8	0,996	0,999	0,994	0,993	0,999	0,998
0,85	0,994	0,993	0,982	0,979	0,994	0,988
0,9	0,991	0,942	0,952	0,901	0,947	0,920
0,95	0,991	0,798	0,864	0,752	0,781	0,751
1	0,981	0,503	0,673	0,534	0,469	0,495
n=200			Population	Biweighted		
Φ	Clásico	Higly(Coman)	Percentage	Midcovarianc	MCD	MVE
0,8	0,998	1,000	0,998	1,000	1,000	1,000
0,85	0,998	1,000	0,998	1,000	1,000	1,000
0,9	0,999	1,000	0,995	0,998	1,000	0,998
0,95	0,999	0,948	0,959	0,930	0,948	0,930
1	0,989	0,509	0,671	0,563	0,465	0,513

Tabla 8.20: Con outlier aditivo y perturbando el espacio muestral. Probabilidad de rechazar la hipótesis nula cuando $\phi \in \{0.8, 0.85, 0.9, 0.95, 1\}$, $\alpha = 0.1$.

Bibliografía

ANDERSON, T.W . (1993). Goodness of fit test for spectral distributions.

ANDERSON, T.W and STEPHENS, M.A. (1993). The modified Cramér-von Mises goodness-of-fit criterion for time series.

ANDERSON, T.W . (1995). Goodness-of-fit test for autorregresive processes.

BHUHLMANN, P., AND KHUNSCH, H. R. (1999). Block length selection in the bootstrap for time series. Computational Statistics and Data Analysis 31 295-310.

BREIMAN, L. (1999). Bagging predictors, Machine learning, 26 Nº2,123-140.

BREIMAN, L. (2001). Random Forest. Machine Learning 45 5-32.

CARLSTEIN, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary time series. Ann. Statist. 14 1171–1179.

CHRISTOPHE CROUX, GENTIANE HAESBROECK. Influence Function and efficiency of the minimium covariance determinant scatter matrix estimator.

DICKEY, D. A., FULLER, W. A., (1981). Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49.

DICKEY, D. A., FULLER, W. A., (1979). Distribution of the estimators for autoregressive time series with unit root. J. Amer. Statist. Assoc. 74.

DE JONG, R. M., C. AMSLER AND P. SCHMIDT. (2007). A Robust Version of the KPSS Test Based on Indicators. Journal of Econometrics 137, 311-333.

DONOHO, D. L. AND HUBER, P. J. (1983) The notion of breakdown point. Wadsworth, Belmont..

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife.Ann. Statist. 7 1–26.

ELLA ROELANT, STEFAN VAN AELST, GERT WILLEMS. The minimum weighted covariance determinant estimator.

ERIC A. CATOR AND HENDRIK P. LOPUHAHA. (2009). Central limit theorem and influence function for the MCD estimators at general multivariate distributions July 1, 2009

GENTON, M. G. AND MA, Y. (1999) Robustness properties of dispersion estimators. Statistics and Probability Letters 44.

GENTON M. G. (2003). Comprehensive definitions of breakdown points for independent and dependent observations. J. R. Statist. Soc. B 65, Part 1, pp. 81–94

HAMPEL, F. R. (1971) A general qualitative definition of robustness. Ann. Math. Stat. 42 1887.

HAMPEL, F. R. (1974) The influence curve and its role in robust estimation. Jour. Am. Stat. Assoc. 69.

HUBER, P. J. (1981) Robust Statistics. New York: Wiley.

KWIATKOWSKI, D., P. C. B. PHILLIPS, P. SCHMIDT AND Y. SHIN. 1992. Testing the Null Hypothesis of Stationarity Against the Alternative of Unit Root. Journal of Econometrics 54, 159-178.

N. A. CAMPBELL1, H. P. LOPUHAA AND P. J. ROUSSEEUW. (1989). On the calculation of a robust S-estimator of a covariance matrix. The Annals of Statistics, Vol 17.

POLITIS, D. N. AND ROMANO, J. P. (1994a). The stationary bootstrap. J. Amer. Statist. Assoc. 89 1303–1313.

LAHIRI. Selecting optimal block lengths for block bootstrap methods S.N

LAHIRI, S. N. (2003). Resampling methods for dependent data. Springer-Verlag, New York, NY.

LOPUHAÄ, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. The Annals of Statistics, 4.

LOPUHAÄ, H. P. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics, Vol. 19.

LOPUHAÄ, H. P. (1992). Highly efficient estimators of multivariate location with breakdown point, The Annals of Statistics 1992, Vol 20, No. 1.

RICARDO MARONNA, DOUG MARTIN AND VICTOR YOHAI. (2006). Robust Statistics. Wiley.

SUN & PANTULA (1999). Testing for trends in correlated data. Statistics and probability letters 41.

VALIANT, L.G. (1984). A theory of the learnable

YANYUAN MA AND MARC G. GENTON. (2002). Highly robust estimation of the autocovariance function.