Hairo Ulises Miranda Belmonte


# Tarea 6



## Ejercicio 1

**Utilice la teoría para modelar y explicar correctamente los resultados obtenidos en su simulación del ejercicio 1 de la sesión 1.**

- **¿Bajo que condiciones es posible recuperar las características del modelo analítico en la simulación?**

Si se pregunta como la simulación se puede aproximar al modelo de Brown, la respuesta sería la siguiente:

Al realizar la simulación y obtener la media de los valores propios muestrales se obtiene una estimación que contempla al sesgo que se hace presenta por un número de histórico de los activos finitos. Si quisiéramos el resultado de Brown, le tendríamos que reducir el sesgo o incrementar el número de muestra de los activos.

Si se pregunta como la simulación se puede aproximar al modelo corregido por Harding, basta con realizar la simulación, ya que la media de los valores propios muestrales, contemplan el sesgo que se genera por muestras finitas.

En resumen , el modelo Brown (ATP) se considera sesgado al estimar con PCA para un solo factor; a menos que se cuente con un histórico mayor para los activos.

Harding, al utilizar la teoría de matrices aleatorias, caracteriza el comportamiento límite de los valores propios muestrales y la distribución del valor propio más grande. Por lo tanto, para obtener una estimación correcta de los factores, se necesita hacer la corrección por el sesgo o una muestra bastante grande.

Lo anterior se muestra en los siguientes gráficos:

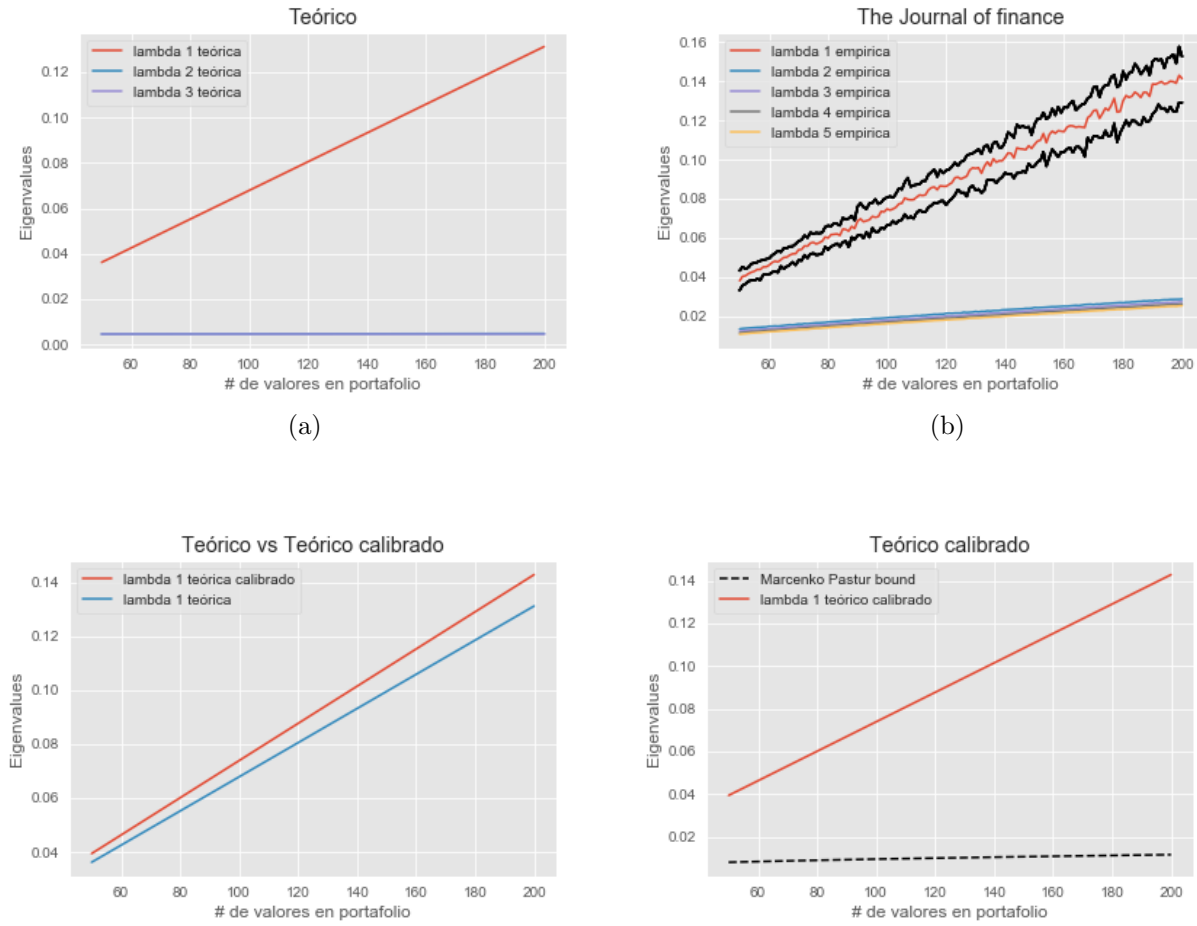
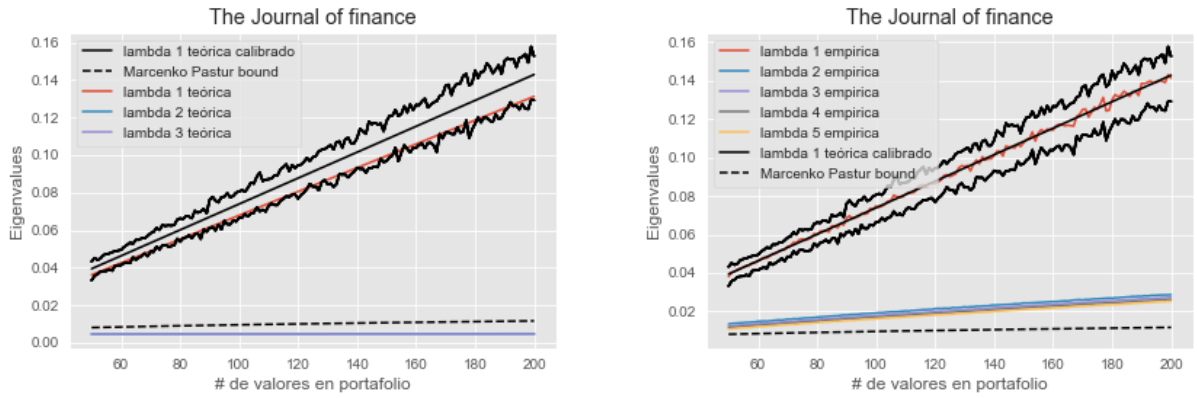



Figure 1: Modelo ATP tradicional y corregido por sesgo

- Figura 1a; presenta los valores propios poblacionales, el resultado de Brown, donde no se considera el sesgo generado por una muestra finita.
- Figura 1b; presenta la simulación de los valores propios muestrales (media) para los primeros cinco valores propios.
- Figura 1c; presenta el valor propio poblacionales más grande, respecto a la expectativa analítica del valor propio más grande de la muestra, que a comparación del primero, sí contemplan el sesgo generado por muestras finitas.
- Figura 1d; presenta a la expectativa analítica del valor propio más grande de la muestra y a la cota de Marcenko - Pastur, esta última delimita a aquellos factores que no aportan información.

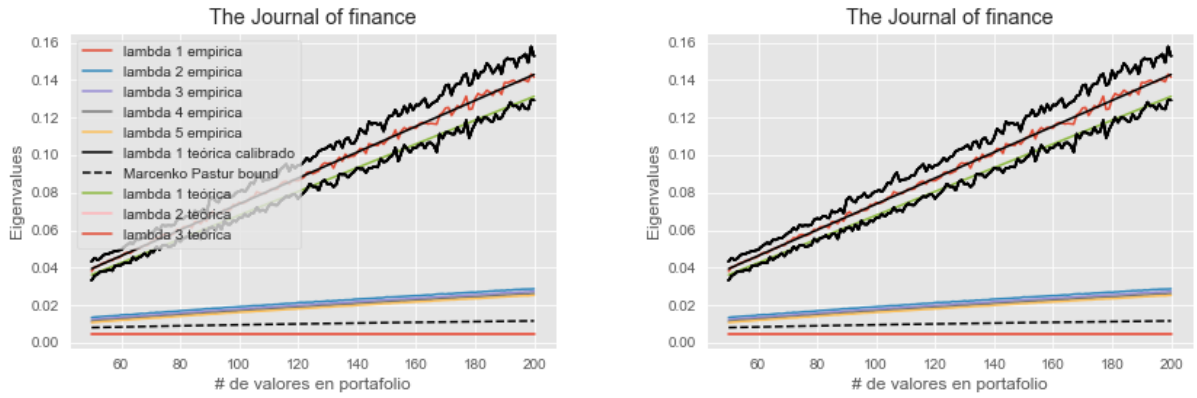



(a) (b)

Figure 2: Modelo ATP tradicional y corregido por sesgo

- Figura 2a; presenta al modelo de Brown, los primeros 3 valores propios poblacionales, respecto al modelo corregido por sesgo de Harding, con la expectativa analítica del valor propio más grande de la muestra. Se observa que el valor propio más grande poblacional, se encuentra el el cuantil inferior del intervalo, de la media del valor propio más grande muestral.

- Figura 2b; presenta la media de los cinco valores propios más grandes (simulación), respecto a la expectativa análitica del valor propio más grande y la cota de Marcenko - Pastur. En este caso se tiene que la media de los valores propios distintos al más grande, se encuentran por arriba de la cota, haciendo énfasis en la relevancia para la estimación de los factores. A su vez, se puede observar como el valor propio más grande de la media muestral, fluctúa sobre el valor más grande ajustado por el sesgo.




(a) (b)

Figure 3: Modelo ATP tradicional y corregido por sesgo

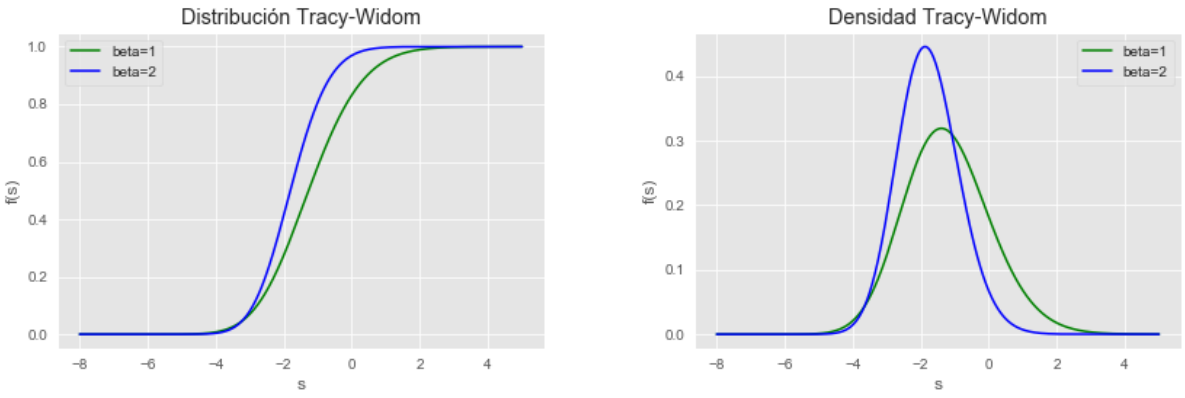
- Figura 3a; presenta al modelo de Brown, los primeros 3 valores propios poblacionales, respecto al modelo corregido por sesgo de Harding, con la expectativa analítica del valor propio más grande de la muestra y la media de los primero cinco valores propios (simulados).
- Figura 3b; se repite la figura 3a pero sin mostrar las etiquetas, con el fin de tener mejor visualización.

En conclusión, se debe tener presente la corrección del sesgo por muestra finita, para calcular los factores del modelo ATP con PCA.

# Ejercicio 2

**a) Revise el ejemplo 6.1.7 de la pag. 170 en Rencher (2002). En este problema se encontró $\theta_{obs} = 0.652$. Aplicando la metodología estudiada en esta sesión encuentre $\theta^{TW}_{0.05}$ y rechace o acepte la hipótesis nula. Discuta los resultados e implicaciones en el contexto del ejemplo estudiado.**

En base a lo realizado de la tarea 5, se presenta la función de densidad y distribución del valor propio más grande, aproximada mediante la distribución Tracy Widom. La figura 4a, muestra la función de distribución a distintos valores de $\beta$; la figura 4b, muestra la función de densidad a distintos valores de $\beta$




(a) (b)

Figure 4: Distribución Traicy Widom

Se realiza el cálculo de $\theta^{TW}$, y se contrasta con el valor propio más grande del test de Roy, y su correspondiente valor crítico - el de tablas- para dar respuesta a la hipótesis de MANOVA. Para realizar el cometido, se utiliza el ajuste correspondiente de la forma:[1]

$$\mathbf{n} = (n - p - 1)/2$$
$$\mathbf{m} = (|n - p| - 1)/2$$
$$N = 2(s + \mathbf{m} + \mathbf{n}) + 1$$

$$\sin^2(\gamma/2) = (s - 0.5)/N$$
$$\sin^2(\phi/2) = (s + 2m + 0.5)/N$$

con:

$$\sigma = (\sigma^3)^{\frac{1}{3}}$$
$$\mu = 2\log(\tan((\phi + \gamma)/2))$$
$$\sigma^3 = (16/N^2)(1/(\sin(\phi + \gamma)^2 \sin(\phi)\sin(\gamma)))$$

obtenemos:[2]

$$f_\alpha = 0.9800000000000857$$

[1]Formulas obtenidas de: Johnstone(2015) APPROXIMATE NULL DISTRIBUTION OF THE LARGEST ROOT IN MULTIVARIATE ANALYSIS.

[2]Nota: se probaron los valores de la lamina para verificar la validez de la tabla $F_1$; se tiene lo siguiente: $f_{0.99} = 2.025$ $f_{0.95} = 0.98$ $f_{0.90} = 0.45$.

$$\theta = exp\{\mu + f_\alpha \sigma\}/(1 + exp\{\mu + f_\alpha \sigma\}) = 0.38390835773363274$$

La tabla 1, muestra los valores de los parámetros que se utilizaron:[3]

Table 1: Parámetros

| **Variable** | Valor |
|---|---|
| $\sigma$ | 0.3026 |
| $\mu$ | -0.7696 |
| $\phi$ | 0.6362 |
| $\gamma$ | 0.5589 |
| $\alpha$ | .05 |
| n | 18.5 |
| m | 0 |
| s | 4 |
| $\theta^{TW}_{0.05}$ | 0.3839 |

Se obtiene un valor de $\theta^{TW}$ muy cercana a lo encontrada con la metodología de la raíz más grande, conocido como: *"Roy's union-intersection test"*, con el valor propio más gande de $\theta = 0.652$, y un valor crítico de 0.377, rechazando la hipótesis nula:

$$\theta^{Obs} = \lambda_1/(1 + \lambda_1) = 0.652$$

De esta manera, se encuentra mediante Traicy-Widom un valor crítico muy cercano al de tablas, con una tasa de error de:

$$r = (\theta^{TW}_\alpha/\theta_\alpha) - 1 = (0.3839/0.377) - 1 = 0.01830$$

**Hipótesis**

En el caso del Test de Roy, se Rechazar $H_o$ si:

$$H_o = \mu_1 = \mu_2 = ... = \mu_k \ si \ \theta > \theta_{\alpha,s,m,N}$$

Nota: cabe mencionar que el $\theta_{\alpha,s,m,N}$, se obtiene mediante tablas con los parámetros mostrados en la tabla 1.

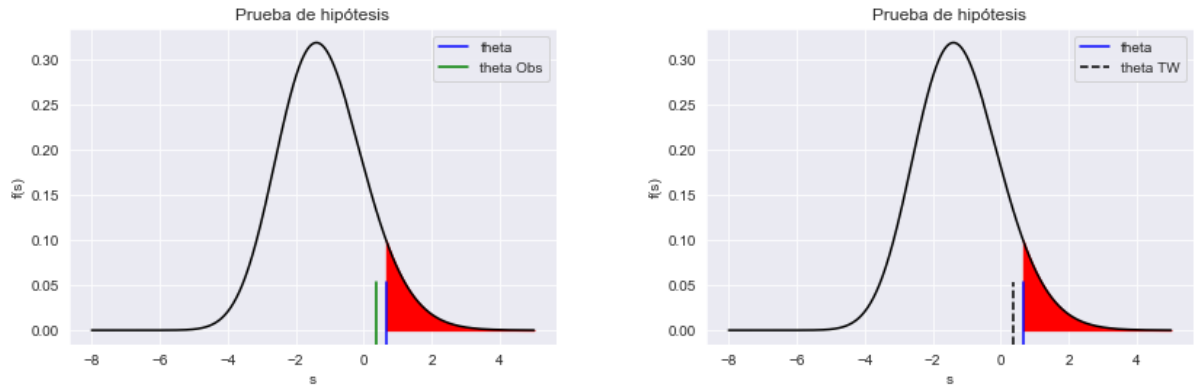
En el caso de $\theta^{TW}$, se rechaza la hipótesis nula si:

[3]Valores que se tomaron del artículo de: Johnstone(2015) APPROXIMATE NULL DISTRIBUTION OF THE LARGEST ROOT IN MULTIVARIATE ANALYSIS.

$$H_o = \mu_1 = \mu_2 = ... = \mu_k \ si \ \theta^{Obs} > \theta_{TW}$$

La figura 5a, muestra la distribución Tracy-Widom, con el estadístico de roy $\theta^{Obs} = 0.652$ (linea azul) contrastado con el valor de tabla $\theta_{\alpha,s,m,N} = 0.377$ (linea verde) al 0.05. Donde se rechaza la hipótesis nula de igualdad de medias en los grupos dado que $\theta^{Obs} > \theta_{\alpha,s,m,N}$.

La figura 5b, muestra la distribución Tracy-Widom, con el estadístico de roy $\theta^{Obs} = 0.652$ (linea azul) contrastado el estimado con Tracy-Widom $\theta^{TW} = 0.3839$ (linea negra). Donde se rechaza la hipótesis nula de igualdad de medias en los grupos dado que $\theta^{Obs} > \theta_{TW}$.




(a) Roy's Test

(b) Aproximación Treacy-Widom

Figure 5: Prueba de hipótesis

En términos de la hipótesis nula, a un nivel de 0.05, se tiene evidencia suficiente para rechazar la hipótesis nula, y por ende la muestra no proviene de la misma población con la misma media dado la estructura de correlación.

**b) Obtenga una muestra de dimensiones pxn, donde p y n son del mismo orden. Rechace o acepte la hipótesis nula de igualdad de varianza utilizando los estadísticos usuales (use y explique al menos dos estadísticos distintos), así como el test de Tracy-widom. Discuta sus resultados**

### Datos

Para el ejercicio se descargan series financieras de yahoo finanzas por medio de la librería *yfinance* de python. Se descargan 84 series financieras (valor más alto del día), con periodicidad diaria, bajo un periodo de 84 días (2018-08-19 a 2018-04-20).[4]

[4]El nombre de las series se ecnuentra en el archivo "Tarea 6-Hairo Ulises-Miranda Belmonte.py", en la

Se utilizan los retornos de las series y se dividen en 4 periodos, de forma aleatoria, sumando a cada periodo una constante $c$ para cada división. Al sumarle un escalar lo que se pretende es realizar un cambio estructural en la serie, y obtener un proceso con 4 medias distintas a través del tiempo.

Lo anterior se hace para obtener una muestra de diferentes grupos y ejemplificar los resultados de cierta manera sabiendo lo que sucederá, ya que con la muestra que procede de diferentes procesos se obtiene el efecto de provenir de distintos grupos, que en el caso de una serie de tiempo se presentan como cambios estructurales sobre los precios de los retornos.

**Pruebas**

Las pruebas que se utilizan son:

- Test de Wilks
- Test de Roy
- Test de Pillar
- Test de Lawley-Hottelling

Los cuatro test buscan probar la Hipótesis nula de iguldad de medias dado la estructura de covarianza.

$$H_o = \mu_1 = \mu_2 = ... = \mu_k$$

A continuación se presentan los estadísticos y una discusión sobre sus ventajas y desventajas.

**Test Wilks**

El estadístico de Wilks se encuentradado por la siguiente expresión:

$$\Lambda = \frac{|E|}{|E + H|}$$

donde la hipótesis nula se rechaza si $\Lambda \leq \Lambda_{\alpha,p,v_H,v_E}$

sección "Ejercicio 2" apartado "b".

Con $H$ como la suma de cuadrados entre los grupos y $H$ la suma de cuadrados dentro de los grupos.

Los parámetros son los siguientes; $p$ es el número de variables; $v_H$ los grados de libertad de la hipótesis; $v_E$ los grados de libertad del error.

**Test de Roy**

Aproximación de las pruebas únion-intersección, basado en el valor propio más grande, siendo la razón del conocido nombre de la prueba "test del valor propio más grande".

El estadístico de Roy tiene la siguiente forma:

$$\theta = \frac{\lambda_1}{1 + \lambda_1}$$

El valor crítico viene dado por $\theta_{\alpha,s,m,N}$, que es determinado por las tablas elaboradas por Pearson and Hartley 1972, Pillai 1964, 1965.

donde la hipótesis nula se rechaza si $\theta^{obs}\theta\alpha, s, m, N$

Los parámetro son los siguientes:

$$s = min(v_H, p)$$

$$m = \frac{1}{2}(|v_H - p| - 1)$$

$$N = \frac{1}{2}(v_E - p - 1)$$

Al igual que los test anteriores, las siguientes dos pruebas se basan en los valores propios de la matriz $E^{-1}H$.

**Test de Pillar**

El estadístico de Pillar se encuentra dado por la expresión:

$$V^{(s)} = Trace[(E + H)^{-1}H] = \Sigma_{i=1}^{s}\frac{\lambda_1}{1 + \lambda_1}$$

Donde se rechaza la hipótesis nula si $V^{(s)} \geq V_{\alpha}^{(s)}$

con los parámetros $s$, $m$ y $N$ definidos en la prueba de Roy.

**Test de Lawley-Hottelling**

El estadístico de Lawley-Hottelling se encuentra dado por la expresión:

$$V^{(s)} = Trace[(E+H)^{-1}] = \Sigma_{i=1}^{s} \lambda_1$$

Donde se rechaza la hipótesis nula si $V^{(s)} \geq V_{\alpha}^{(s)}$

rechazando la hipótesis nula para un valor grande del estadístico.

**Diferencias**

Cuando la hipótesis de grados de libertad, $v_H$, es uno, las cuatro estadísticas de prueba conducirán a resultados idénticos. Cuando $v_H > 1$, las cuatro estadísticas generalmente conducirán al mismo resultado.

La traza de la Lambda de Wilks, Lawley-Hottelling y la raíz más grande de Roy son a menudo más poderosas que la traza de Pillai si $v_H > 1$. La traza de Pillai es más robusta a las desviaciones de los supuestos que los otros tres (Johnson, R. y Wichern, D., (2014). Applied Multivariate Statistic Analysis. Capítulo 7., pp. 270-275).

**Resultados**

Se utiliza la función *manova* de la librería *statsmodels.multivariate* de python. En base a la función se obtienen la prueba de hipótesis de igualdad en las medias dado la covarianza mediante los estadísticos mencionados. En la siguiente tabla se presentan los resultados.

Table 2: Prueba de Hipótesis MANOVA

| **y** | Valor | p-value |
|---|---|---|
| Wilks' lambda | 0.0000 | 0.0449 |
| Pillai's trace | 1.0000 | 0.0449 |
| Hotelling-Lawley trace | 26356.7431 | 0.2199 |
| Roy's greatest root | 26356.7431 | 0.0449 |

**Conclusión**

En base al p-valor, contrastando a un nivel $\alpha = 0.05$, el estadístico de Wilks, Roy y Roys, rechazan la hipótesis nula, pero de una manera no muy contundente, si se prueba con $\alpha = 0.01$, no se presenta evidencia suficiente para rechazar la hipótesis nula. En el caso del estadístico de Hotellin-Lawley, para un $\alpha = 0.1$ , 0.05 y 0.01, no se tiene evidencia suficiente para rechazar la hipótesis nula. En general, se puede decir, que para ciertos estadísticos y nivel de significancia los grupos tienen la misma media dado la estructura de correlación.

**Traicy-Widom**

Se implementa la aproximación de Traicy-Widom con los siguientes valores de los parámetros:

Table 3: Parámetros

| **Variable** | Valor |
|---|---|
| $\sigma$ | 0.0486 |
| $\mu$ | 0.2088 |
| $\phi$ | 0.8375 |
| $\gamma$ | 0.8375 |
| $\alpha$ | .95 |
| n | -0.5 |
| m | -0.5 |
| s | 84 |
| $\theta_{0.05}^{TW}$ | 0.5637 |

La hipótesis nula es la misma que los estadísticos anteriores. En la tabla 3, se observa el valor crítico estimado con Traicy Widom, $\theta_{0.05}^{TW} = 0.5637$. Se utiliza la distribución de Traicy-Widom ($F_1$), y en base al valor crítico se calcula el área bajo la curva para obtener su p-valor, el cual registra un valor de 0.0867, que es ligeramente mayor a $\alpha = 0.05$. Por lo tanto, dado a ese nivel de significancia, no se tiene evidencia para rechazar la hipótesis nula, por ende, los grupos provienen de la misma distribución con la misma media.

# Ejercicio 3

**Reproduzca la figura anterior implementando el código en Python. Describa las funciones y librerías que utilizó**

Se calcula la suma de densidades espectrales media de dos matrices aleatorias. $M$, una matriz aleatoria de ensamble GOE, con sus elementos que distribuyen $X \sim N(0,1)$, y una matriz $Y$, con elementos reales y distribuye $W \sim W(0,I)$, sin estructura de correlación.

Se realiza mil simulaciones de la suma pesadas por cierta "p", que determina la forma de la densidad de la transformada $H$.

$$M = XX'/2$$
$$Y = WW'$$
$$H = pM/\sqrt{(n)} + (1-p)Y/n$$

- Se almacenan valores propios de H[5]
- Se gráfica el histograma

**Librerías:**

Se utiliza la biblioteca *sympy* para la solución simbólica, en especifico los siguientes módulos:

- solve: resuelve ecuaciones o sistemas
- symbols: genera objeto para operaciones futuras
- im: obtiene la parte imaginaria de algún arreglo

**Parámetros:**

Table 4: Parámetros de la simulación

| **Descripción** | Símbolo | Valor |
|---|---|---|
| Probabilidad | p | 0.3, 0.5, 0.7 |
| Espacio Fila | n | 100 |
| Iteraciones | t | 1000 |
| Espacio columna | T | 150 |

**Solución Simbólica:**

Se resuelve la siguiente ecuación para G:

$$(w^2)(G/2) + ((1-w)(1+a))/(1-(1-w)G) + (1/G) - z = 0$$

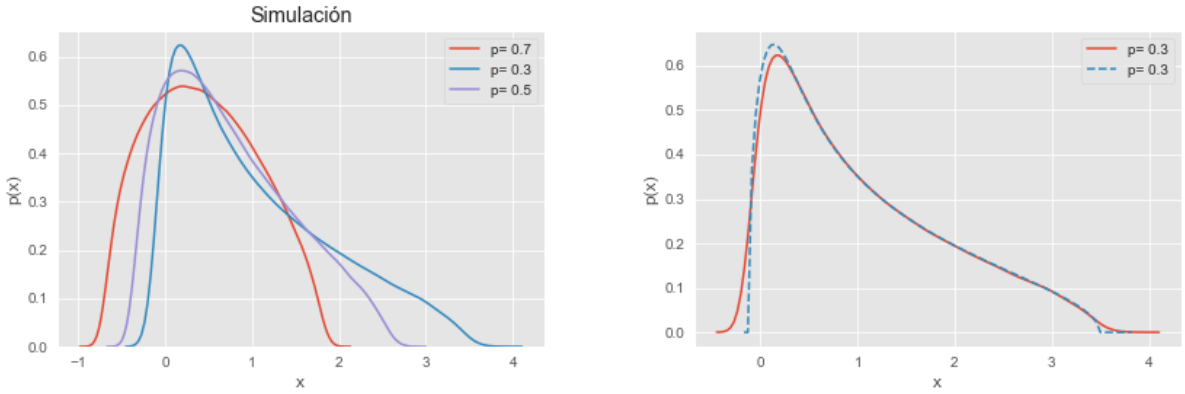
[5]Nota, el ensamble GOE se multiplica por un factor $1/\sqrt{(\beta n)}$ con $\beta = 1$

Table 5: Parámetros de solución simbólica

| **Descripción** | Símbolo | Valor |
|---|---|---|
| Variable | $\alpha$ | (1-c)/c |
| Constante | c | n/T |
| Iteración | Ntp | 100 |
| Espacio Fila | n | 100 |
| Espacio columna | T | 150 |
| Probabilidad | p | 0.3, 0.5, 0.7 |
| Grid | x | min(Simulación) a max(Simulación) |

Por medio de la función *solve* encontramos la soluciones[6]. La solución es un vector de tres dimensiones cuyos componentes representan una solución polinomial de grado tres para G. Se sustituyen los valores $\alpha$ y $p$, y sustituimos para z los valores definidos en el grid x.

**Visualización:**

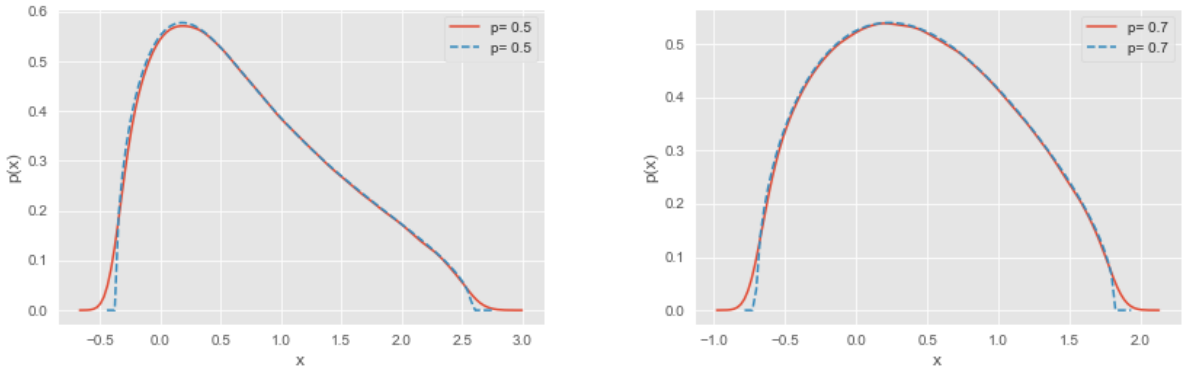



(a) (b)

Figure 6: Transformada

- Figura 6a; muestra la simulación de la transformada; con $p = 0.3$ la forma de la densidad es la de una Wishart; con $p = 0.5$, la forma de la densidad parece mezcla de wishart y Goe; con $p = 0.7$ la forma de la densidad es la de un ensamble GOE
- Figura 6b; muestra la simulación y solución simbólica para $p = 0.3$

[6]La solución de las ecuaciones se encuentran en el archivo "Tarea 6-TW-Hairo Ulises-Miranda Belmonte.py", Sección Ejercicio 4.




(a) (b)

Figure 7: Transformada

- Figura 7a; muestra la simulación y solución simbólica para $p = 0.5$
- Figura 7b; muestra la simulación y solución simbólica para $p = 0.7$

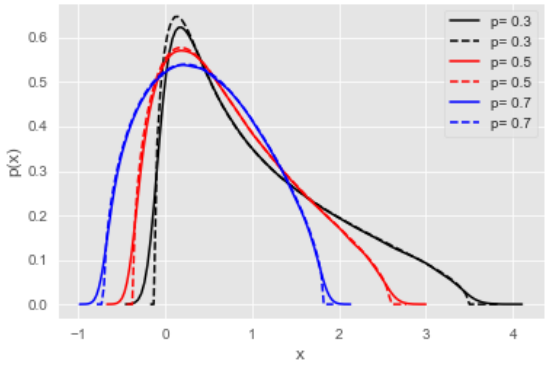



Figure 8: Transformada

- Figura 8; muestra la simulación y solución simbólica para distintos valores de $p$

En conclusión, se llega a la misma forma de la densidad de la transformada al utilizar simulación como la solución numérica.

# Ejercicio 4

Lea el artículo "Nonlinear random matrix theory for deep learing, Pennington and Wirah, 2017"; y escriba un ensayo de 500 a 1000 palabras donde se discutan las siguientes preguntas:

- ¿ Cuál es la contribución más importante del artículo en el contexto de deep learning?
- ¿ Cuál es la contribución más importante del artículo en el contexto de matrices aleatorias?
- ¿ Cuál es la contribución práctica del artículo?
- ¿ Cuales son los alcances y limitaciones de la propuesta del artículo?
- ¿ Considera relevante implementar esta propuesta?

La contribución más importante en el contexto de deep learning es la propuesta de una nueva función de activación, en la cual el proceso de aprendizaje no se interrumpe por problemas de saturación.

En contraste a mecanismos de regularización y prevención de la saturación, el método *batch normalization*, lidia con cambios en la estructura de covarianza de los datos de entrada; no obstante, dentro de la contribución en el artículo se encuentra que *batch notmal ization* se encarga de los primeros momentos, siendo incapaz de atender momentos en la distribución de los datos mayores al los primeros dos.

Es aquí donde entra una de las contribuciones más relevantes, porque al aproximar el problema no lineal por medio de matrices aleatorias, en especifico, el método de momentos, se logra obtener expresiones capaces de controlar momentos mayores, dando estabilidad a los valores singulares a la distribución de la capa de salida de una red neuronal.

Entre la relevancia de la función de activación, radica el resultado de que al tener una variable de salida, (i.e., después de realizar el producto punto de los pesos y las variables de entrada, con una mapeo elemento a elementos de una función no lineal), bajo ciertas condiciones, se obtiene una función de activación capaz de propagar la función de distribución espectral de los elementos de entrada a la variable de salida, implicando que la función de distribución espectral queda intacta en el proceso de entrenamiento de la red neurona, y por ende, evitando problemas de saturación.

Entre las contribuciones para el área de matrices aleatorias se encuentra el estudio respecto a la no linealidad y la matriz de Gram - su distribución espectral-, dando resultados métodos más robustos para problemas de saturación, utilizando resultados conocidos en materia de matrices aleatorias y mejorando los métodos de entrenamiento en deep learning.

La contribución practica del estudio se presenta en la propuesta de una nueva capa de activación en la cual permite heredar la densidad espectral bajo ciertas condiciones, permite agilizar el tiempo de entrenamiento al hacer un ajuste en los pesos sin presentar problemas de saturación; a su vez, mejora el problema de cambio de covarianza, no solo como lo hace *batch normalization*, sino que asegura que la distribución espectral sea la misma a la los datos - bajo ciertas condiciones que menciona el artículo-, implementando un método que generaliza de mejor manera y homogeneizando de cierta forma la distribución de la muestra.

Entre los alcances del artículo se encuentra la mejora en el entrenamiento por medio de una capa de activación que permite estandarizar en cierto aspecto los datos de salida de las capas ocultas. Entre las limitantes se observa que el caso se centra solo en la capa de salida, pero al tener una función de activación la cual implica el calculo de los resolventes, la estimación de los graidientes en el entrenamiento, en mi perspectiva, se vuelve más complejo, que encontraste a utilizar una función de activación como la *relu* (valores negativos los hace cero), o alguna variante.

Entonces, aquí entraría en conflicto en saber si el método propuesto en realidad permite un calculo más eficiente en arquitecturas pequeñas, donde no se necesita mucho tiempo de entrenamiento y no se presentan alteraciones tan pronunciadas en la escala de los datos, permitiendo una propagación adecuada de la información sin presentar cambios en la estructura de covarianza.

El método propuesto es bastante interesante, y relevante en ciertos casos, el simple hecho de presenta datos de entrenamiento con cambios en la estructura de covarianza, en donde el algoritmo al ser entrenado con datos de alguna distribución y probado con datos de alguna otra muestra, tiende a presentar desempeños inadecuados, pero bajo la propuesta del artículo se puede lidiar con esto de una manera más estable que el método de *batch normalization*, dando como resultado modelos con mejor capacidad de generalizar.