Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Matrices aleatorias: 6

Jorge Luis Ramos Zavaleta

6 de junio de 2019

1. Ejercicio

En este ejercicio se busca integrar los conocimientos aprendidos a lo largo del curso. Para ello se solicita realizar lo siguiente:

1. Completar la derivación de la distribución de Marcenko-Pastur, partiendo de las notas de clase. Sea los más claro posible, sin omitir ningún detalle algebraico (puede escanearlo.)

2. Reproduzca la figura 14.1 del libro seguido en este módulo (Introduction to Random Matrices, G. Livan et. al.), bajo las mismas condiciones y parámetros (compruebe que $p > 0{,}05$ en el test de Kolmogorov-Smirnov).

3. Descargue las series de tiempo que componen el ındice bursátil Standard Poor's 500. Utilizando una periodicidad semanal durante los últimos 10 años (Enero 2008 a la fecha).

4. Aplique las transformaciones necesarias (aprendidas en el módulo de series de tiempo) para trabajar las series de tiempo desde el punto de vista estacionario. Deseche las series de los mercados que presentan problemas.

5. Determine el número de componentes significativos adecuando un test derivado de la distribución de Marcenko-Pastur.

6. Aplique regresión por componentes principales utilizando el número de componentes sugeridos por el resultado de matrices aleatorias y compare el resultado utilizando el criterio del 80 % de la varianza. Se busca predecir el valor de apertura del ındice SP 500 el lunes por la mañana a través de los 500 mercados que lo componen.

7. Grafique la efectividad del pronóstico durante este año.

8. ¿Cómo mejoraria el pronóstico? si obtiene un promedio en la efectividad mayor al 50 gana puntos extras en proporción a como este valor se acerque al 100 otros métodos de pronóstico en busca de mayor efectividad, pero siempre contrastando con el criterio de matrices aleatorias).

1.1. Solución

El primer inciso se adjunta al final del documento.

Para el segundo inciso al reproducir el experimento se obtuvo un p-value de 0.3079 usando la prueba de Kolmogorov-Smirnov.

```
Two-sample Kolmogorov-Smirnov test

data:  L/200 and temp
D = 0.013734, p-value = 0.3079
alternative hypothesis: two-sided
```

Sin embargo, varios expertos en el tema [1]indican que algunas pruebas de normalidad como la de Kolmogorov-Smirnov pueden no ser confiables cuando el tamaño de la muestra es muy grande por lo que no es recomendable su uso en solitario con muestras de tamaño mayor a 200 ya que pueden indicar un resultado incorrecto. Por lo que se recomienda hacer uso de criterios graficos o de pruebas mas actualizadas, en este caso se realizo un qqplot de las dos distribuciones. En la figura 1.1 se puede apreciar dicho qqplot y se puede verificar que en general se forma una recta dando un poco mas de certeza de que la distribución generada proviene de la teórica. Aparte se dibujaron las 2 distribuciones en la figura 1.2 para verificar visualmente que se parecen.

Por último se procedio a generar la grafica que contiene el libro, mostrando que la distribución de Marcenco-Pastur es independiente de la elección del β. En la figura 1.5 se puede ver dicha grafica.

[1] https://www.researchgate.net/post/
/Which_statistical_methods_should_be_used_to_test_the_distribution_of_a_small_o_large_sample
https://www.researchgate.net/post/Whats_the_difference_between_Kolmogorov-Smirnov_test_and_Shapiro-Wilk_test

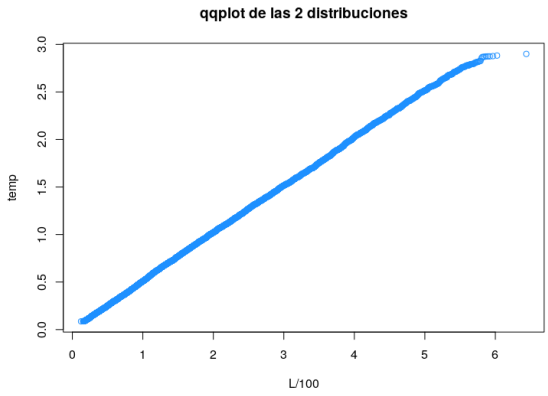

Figura 1.1: Qqplot de las 2 distribuciones

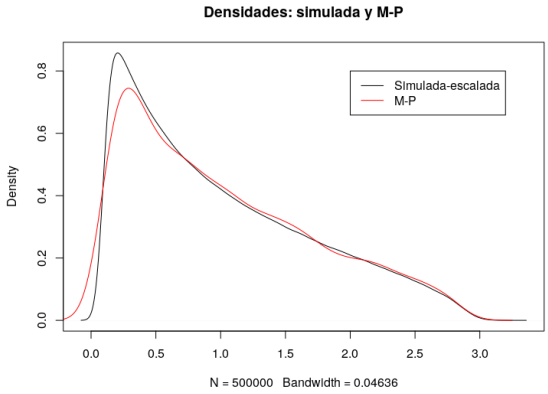

Figura 1.2: Distribuciones simulada y teorica

1.2. S&P 500

Para la siguiente parte del ejercicio se consideraron solo las componentes actuales que integran el indice de S&P, ya con la apropiada limpieza debido a que no todas las componentes estaban presentes durante el periodo que se pide y algunas no contienen todos los datos. Después se realizo una prueba de estacionariedad para cada una de las series encontrando que tanto algunas series de los componentes como el indicador global son no estacionarias, pero al considerar la primera diferencia la prueba de estacionariedad nos indica que todas las series son estacionarias bajo esta transformación al 95 %.

Debido a que los valores de los componentes se encuentran en la misma escala no se encuentra necesario realizar ninguna otra transformación a los datos de las series. Ya con

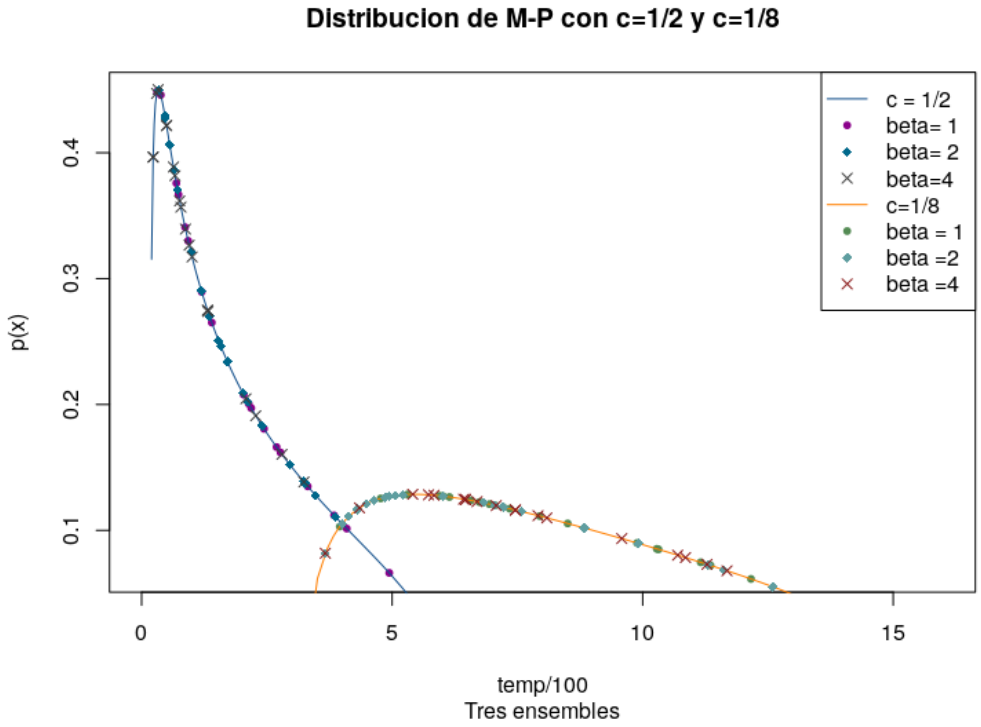

Figura 1.3: Replica de la grafica del libro

el preprocesado y eliminando las series que no nos servian nos quedamos con 450 series de las 500. Aplicando el criterio de Marcenco-Pastur tenemos un valor cercano a 4.41 por lo que se decidieron usar 5 componentes bajo este criterio y bajo el criterio del 80 % de la varianza como se puede observar en la figura se decidio usar 19 componentes principales. Para generar las regresiones solo se consideraron los datos hasta finales del 2017, y se predijeron los correspondientes los valores de apertura para los lunes del 2018 hasta el 19 de Noviembre de 2018, los resultados en diferencias se muestran en las figuras 1.6 y 1.7 donde puede apreciarse cambios muy pequeños. Al regresarlo a sus valores en niveles pueden apreciarse diferencias mas amplias como se puede ver en la figura 1.8.

Para verificar la eficiencia del modelo se grafico la efectividad de la prediccion (1-Error relativo) para cada valor de 2018. Como se observa en la figura 1.9 se genera una efectividad muy alta para cada uno de los valores predichos con respecto de los actuales. En particular para el lunes 19 de Noviembre de 2018 se predijo un valor del indice de 2657.658 y tenemos un valor real de 2730.74 y de acuerdo al criterio de efectividad tenemos

$$Efectividad(19-11-2018) = 1 - \frac{|2657{,}658 - 2730{,}74|}{2730{,}74} = 1 - \frac{73{,}082}{2730{,}74} = 1 - 0{,}027 = 0{,}973$$

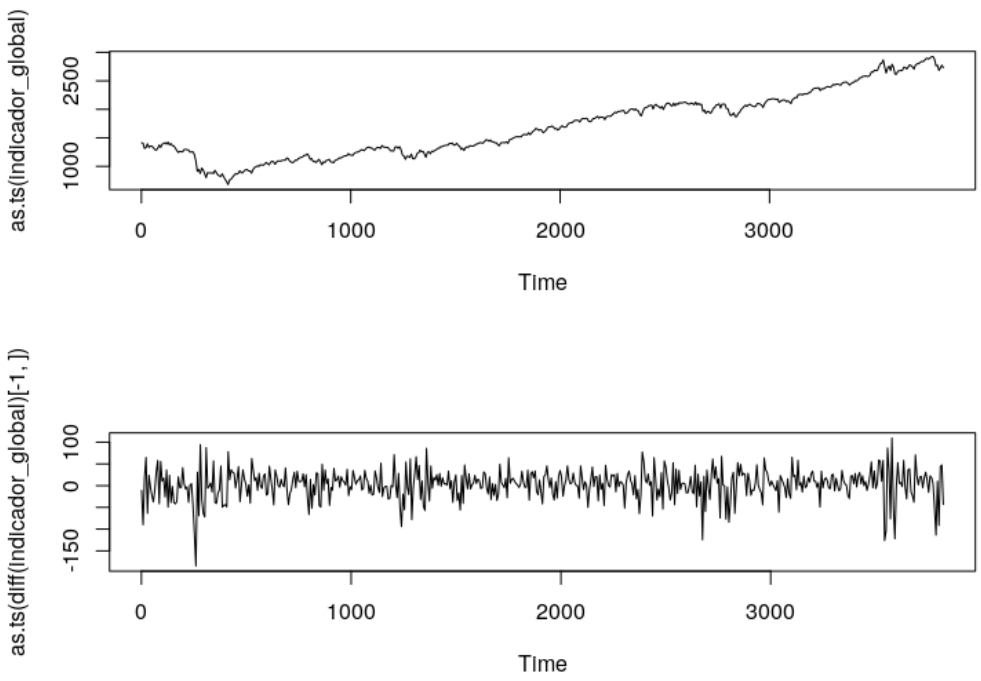

Figura 1.4: Series del indicador en valores en niveles y en la primera diferencia. Se puede observar que con la primera diferencia parece ser estacionaria.

por lo que tendriamos una efectividad del 97.3 % aproximadamente y en promedio para 2018 se obtuve un 97.99278 % de efectividad. Se genero otro criterio para verificar estos resultados ya que la diferencia entre el valor predicho y el real es de 73 puntos que en el mundo financiero es muy alta, para ello verificabamos el cambio de signo de un lunes al otro, es decir verificamos si el indice aumentaba o disminuia y comparabamos dichos cambios entre los valores actuales y de los valores actuales contra los predichos. Usando este criterio se encontro un 62.22 % de precision, es decir se identificaban las subidas y bajadas del indice un 62 % de las veces.

Una de las cosas que se pueden hacer para mejorar la predicción es usar mas componentes, sin embargo esto puede agregar ruido innecesario a los datos y puede terminar generando un sobreajuste del modelo. Otra opción es elegir series mas cortas donde las series que componen actualmente el indice esten todas completas, ya que al no encontrarse completas desde estas fechas algunas series se desecharon y entre ellas pudo haberse quitado una de las series que pudiera tener un mayor impacto en el indice.

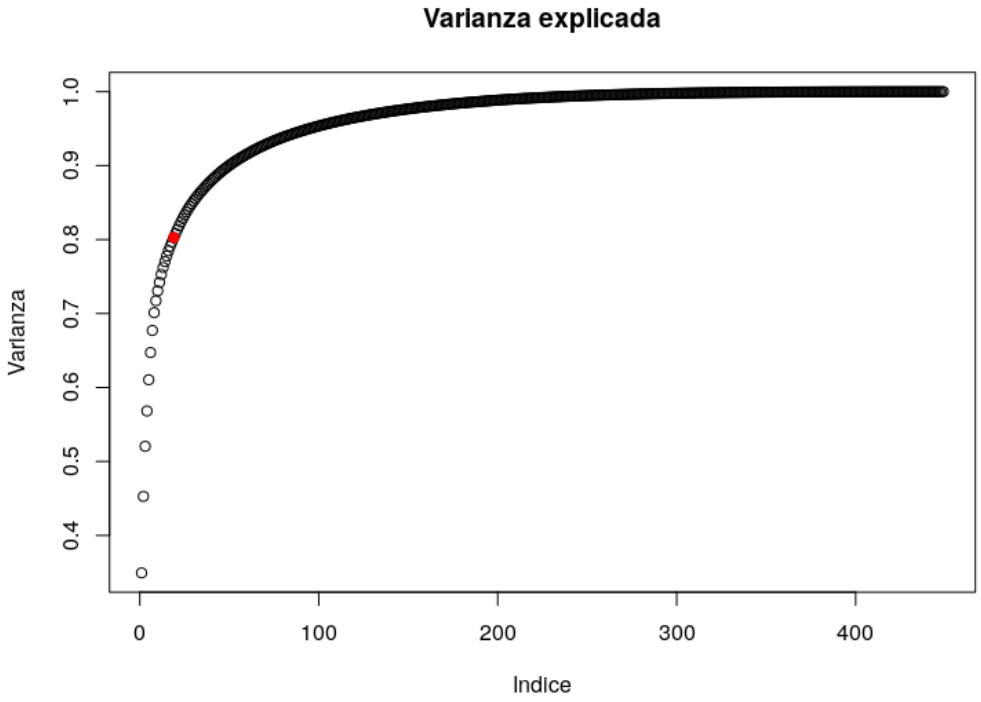

Figura 1.5: Varianza explicada por los componentes principales. En rojo la varianza explicada por el componente 19

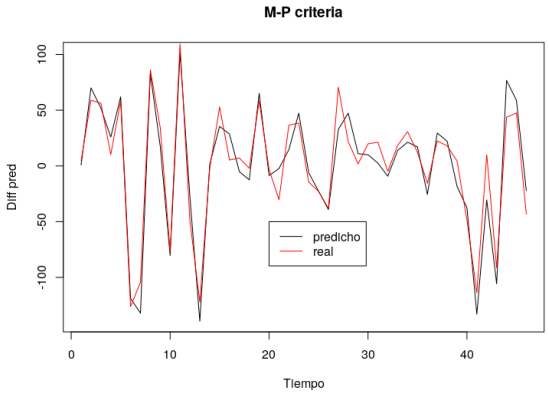

Figura 1.6: Predicho vs real para el caso del criterio de Marcenco-Pastur

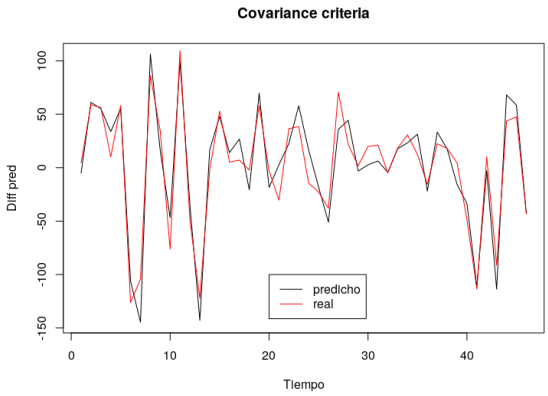

Figura 1.7: Predicho vs real para el caso del criterio de 80 % de varianza

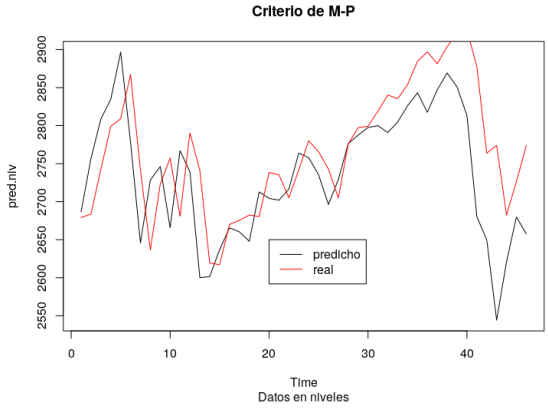

Figura 1.8: Predicho vs real para el caso del criterio de Marcenco-Pastur con los datos en niveles

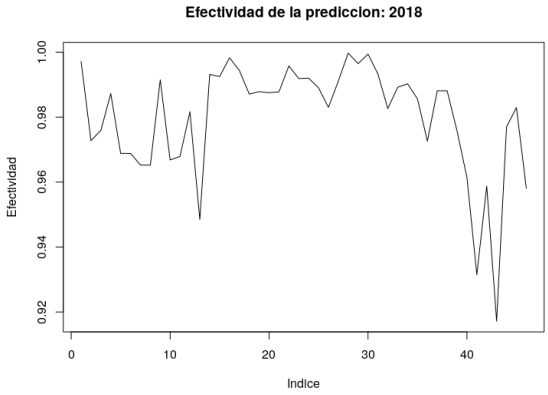

Figura 1.9: Efectividad del pronostico