Matrices aleatorias (Tarea1)

J. Antonio García Ramirez

5 de semtiembre 2018

Ejercicio 1. De las diapositivas correspondientes a esta sesión, completar los pasos para llegar de la ec. 6 a la ec. 9.

Me queda claro, intuitivamente, la idea de conservación de las partículas sobre un eje independientemente del tiempo, lo considero como una v.a. definida sobre todo R (por eso su integral es finita y la podemos normalizar) entonces partimos del supuesto (o ley) de la conservación de las partículas:

$$f(x, t+\tau) = \int_{-\infty}^{\infty} d\Delta\rho(\Delta) f(x-\Delta, t)$$

Como la función de densidad f está definida sobre R^2 podemos aproximar la función alrededor de un punto $(x, t+\tau)$ con la serie de Taylor, en particular podemos aproximar la función de un punto en cada dirección por lo que de la ecuación anterior del lado izquierdo consideramos su expansión en Taylor en la dirección $t+\tau$ y del lado derecho la expansión en la dirección $x-\Delta$, teniendo:

$$\begin{aligned} f(x,t) + \tau\frac{df}{dt} + ... &= \int_{-\infty}^{\infty} d\Delta\rho(\Delta)\left[f(x,t) - \Delta\frac{df}{dx} + \frac{\Delta^2}{2}\frac{d^2f}{dx^2} + ...\right] \\ &= f(x,t)\int_{-\infty}^{\infty} d\Delta\rho(\Delta) - \frac{df}{dx}\int_{-\infty}^{\infty}\Delta d\Delta\rho(\Delta) + \frac{1}{2}\frac{d^2f}{dx^2}\int_{-\infty}^{\infty}\Delta^2 d\Delta\rho(\Delta) + ... \end{aligned}$$

Como $d\Delta p(\Delta)$ es una función de densidad su integral sobre todo R vale la unidad por lo que el segundo término del lado derecho es la derivada de una constante. Además como asumimos que la función de densidad $d\Delta p(\Delta)$ es simétrica el tercer término de la ecuación anterior vale cero (al igual que todas las integrales de $\Delta^n d\Delta p(\Delta)$ con n inpar) entonces:

$$f(x,t) + \tau\frac{df}{dt} + ... = f(x,t) + \frac{1}{2}\frac{d^2f}{dx^2}\int_{-\infty}^{\infty}\Delta^2 d\Delta\rho(\Delta) + ...$$

Entonces es razonable truncar la serie de Taylor (de ambos lados hasta el segundo orden) por lo que tenemos:

$$f(x,t) + \tau\frac{df}{dt} + R_2 = f(x,t) + \frac{1}{2}\frac{d^2f}{dx^2}\int_{-\infty}^{\infty}\Delta^2 d\Delta\rho(\Delta) + R_2^*$$

Donde las $R's$ denotan los residuos correspondientes que consideramos despreciables. Luego despejamos la parcial con respecto al tiempo

$$\frac{df}{dt} = D\frac{d^2f}{dx^2}$$

Donde la constante D es la misma que definimos en clase y por ser una integral de una función de densidad es finita.

La ecuación anterior se reconoce como la ecuación de difusión del calor (en una vara de ancho muy pequeño), cuya transformada de Fourier es:

$$f(x,t) = \int_{-\infty}^{\infty} e^{-2\pi i y x} f(y,t) dy$$

Y recordando el teorema fundamental del cálculo (que se extiende a funciones analiticas) y derivando la ultima expresión con respecto a x dos veces tenemos:

$$\frac{df}{dt} = \frac{d^2f}{dx^2}\int_{-\infty}^{\infty}\frac{\Delta^2}{2\tau}d\Delta\rho(\Delta) = D\frac{d^2f}{dx^2}$$

Y sustituyendo en la antepasada ecuación pasamos de una ecuación parcial a una ordinaria :D, teniendo:

$$\frac{d^2f}{dx^2} = (2\pi ix)^2 f(x,t) = -4\pi^2x^2f$$

Y entonces:

$$\frac{df}{dt} = D(-4\pi^2x^2f)$$

Cuya solución es:

$$f(x,t) = \int_{-\infty}^{\infty} e^{-2\pi iyx}f(y,t)dy$$

Ahora viene lo interesante, la condición inicial $f(x,0) = \delta(x)$ lo que permite la unanicidad de la solución, digo que es un condición interesante por que es la primera vez que veo a la delta de dirac tan tangible pues lo interpretó como una densidad en un el tiempo cero es decir que en el tiempo cero se tienen todas las partículas lo cual intuitivamente es evidente. La inversa de la transformada de fourier (cuya prueba de existencia sirve para garantizar la caracterización de v.a. con sus funciones generadas de momentos) de la delta de direc es $

$\delta(x) = \int_{-\infty}^{\infty}\delta(y)e^{-i\omega y}dy = 1$

Sustituyendo tenemos que

$$f = e^{-4\pi^2Dx^2t}$$

Cuya inversa de Fourier es:

$$f(x,t) = \frac{1}{\sqrt{4\pi Dt}}e^{\frac{-x^2}{4Dt}}$$

Verificaquemos que en verdad tenemos una solución:

Derivando con respecto al tiempo $f(x,t)$ tenemos

$$\frac{df}{dt} = -\frac{e^{-\frac{x^2}{4Dt}}}{4\sqrt{\pi D}t^{3/2}} + \frac{x^2e^{-\frac{x^2}{4Dt}}}{8\sqrt{\pi}D^{3/2}t^{5/2}}$$

Y derivando dos veces con respecto a x tenemos:

$$D\frac{d^2f}{dx^2} = -\frac{e^{-\frac{x^2}{4Dt}}}{4\sqrt{\pi D}t^{3/2}} + \left(-\frac{x}{4\sqrt{\pi}(Dt)^{3/2}}\right)\left(-\frac{x}{2t}e^{-\frac{x^2}{4Dt}}\right) = -\frac{e^{-\frac{x^2}{4Dt}}}{4\sqrt{\pi D}t^{3/2}} + \frac{x^2e^{-\frac{x^2}{4Dt}}}{8\sqrt{\pi}D^{3/2}t^{5/2}}$$

Ejercicio 2. A través de una simulación numérica, mostrar la convergencia en distribución de la suma de $n \in \{2, 3, 1000\}$ variables aleatorias i.i.d., para los procesos estocásticos con la siguiente función de densidad:

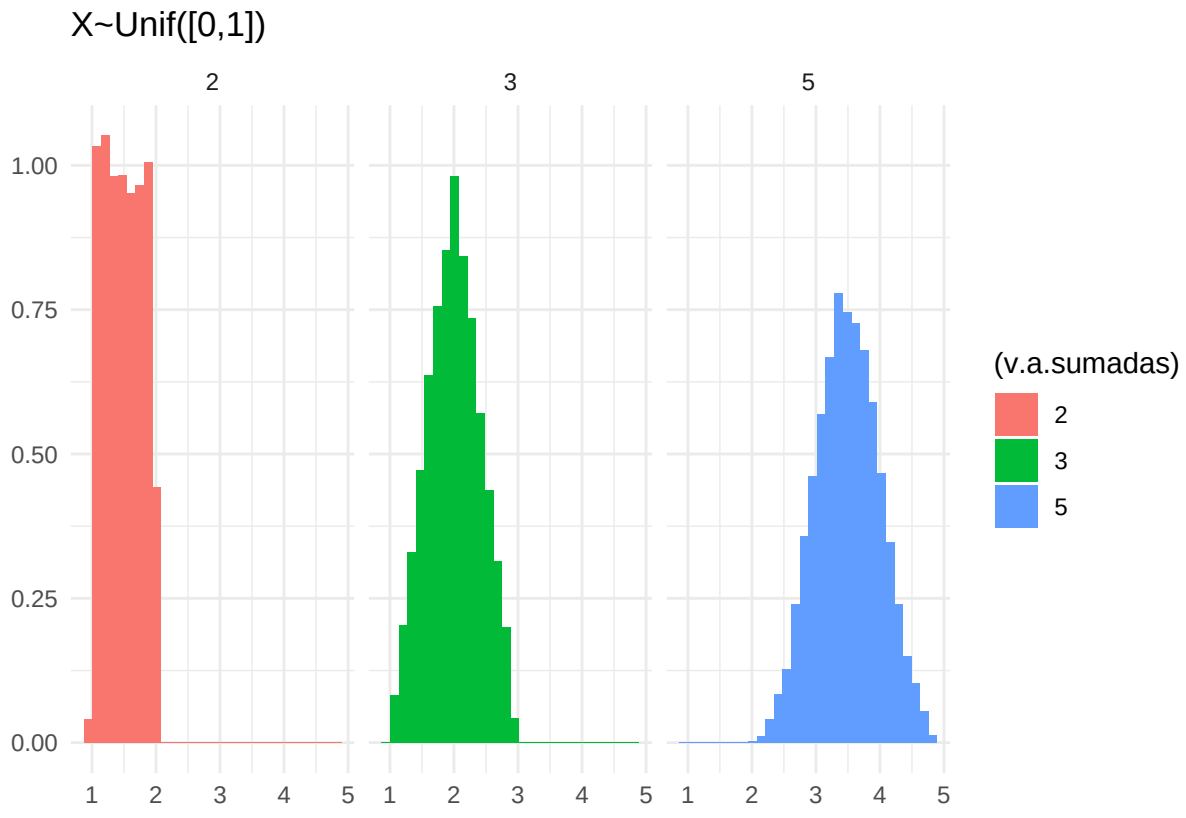
Construiré una función para realizar las simulaciones.

Después de realizar las simulaciones para la v.a. $X_i \sim Unif([0,1]), Z_i \sim Exp(\lambda = 1), Y_i \sim N(0,1)$ se logra notar que las suma de v.a.´s no es estable y la suma de $n \in \{2, 3, 1000\}$ converge a una normal.

```
library(ggplot2)
library(reshape2)
Simula <- function(distribucion, no.suma, parametros, N)
{
  simple.path <- mapply(function(x){
    muestra <- distribucion(x,parametros)
    medias <- c(sum(muestra), x)
    names(medias) <- c('realizacion', 'v.a.sumadas')
    return(medias)   },no.suma  )
}
```

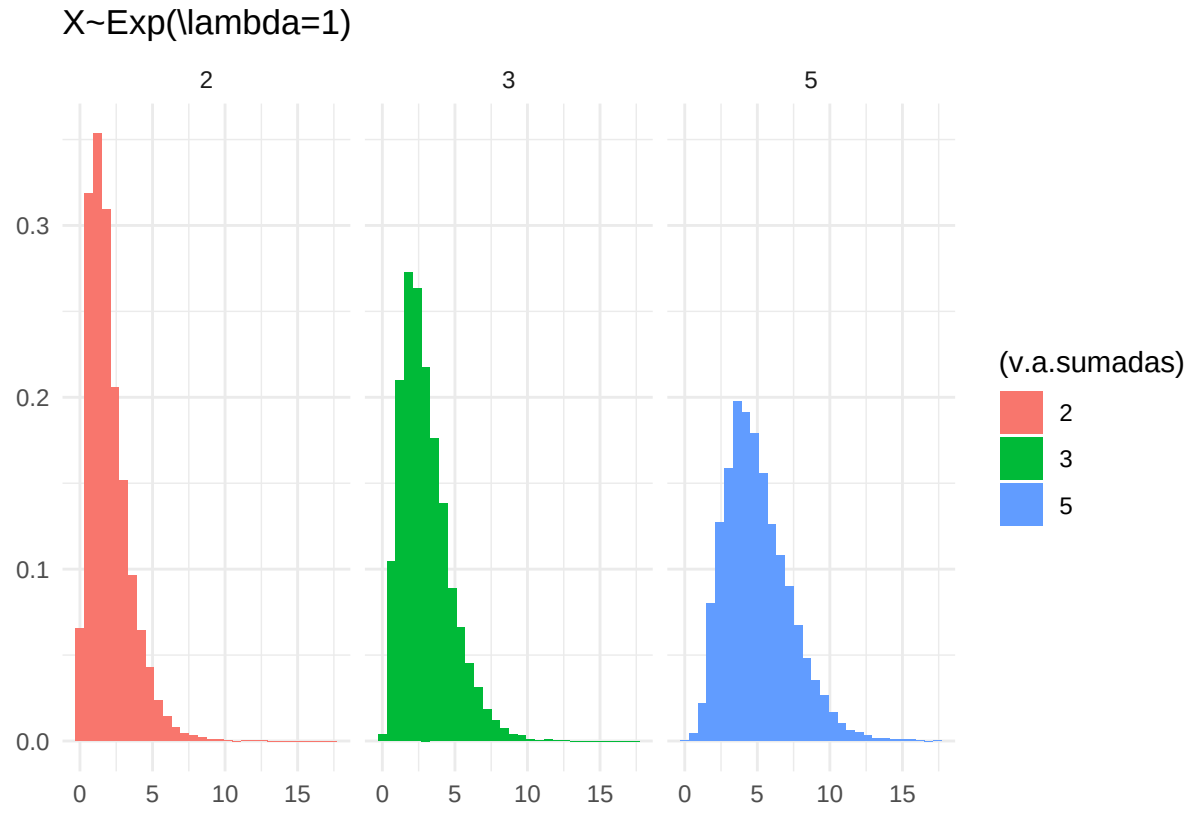
i. $X_i \sim Unif([0,1])$

```
set.seed(123)
repeticiones <- 10000
ensamble.normal <- lapply(1:repeticiones,FUN=Simula,distribucion=runif,
                          no.suma=c(2, 3, 5), parametros=c(0,1))
ensamble.normal <- as.data.frame(t(as.data.frame(ensamble.normal)))
ensamble.normal$v.a.sumadas <- factor(ensamble.normal$v.a.sumadas)
ggplot(ensamble.normal, aes(x=realizacion, fill=(v.a.sumadas)))+
  geom_histogram(aes(y=..density..))+ facet_wrap(~v.a.sumadas, ncol=3) +
  theme_minimal() +ggtitle('X~Unif([0,1])') +xlab('')+ylab('')
```



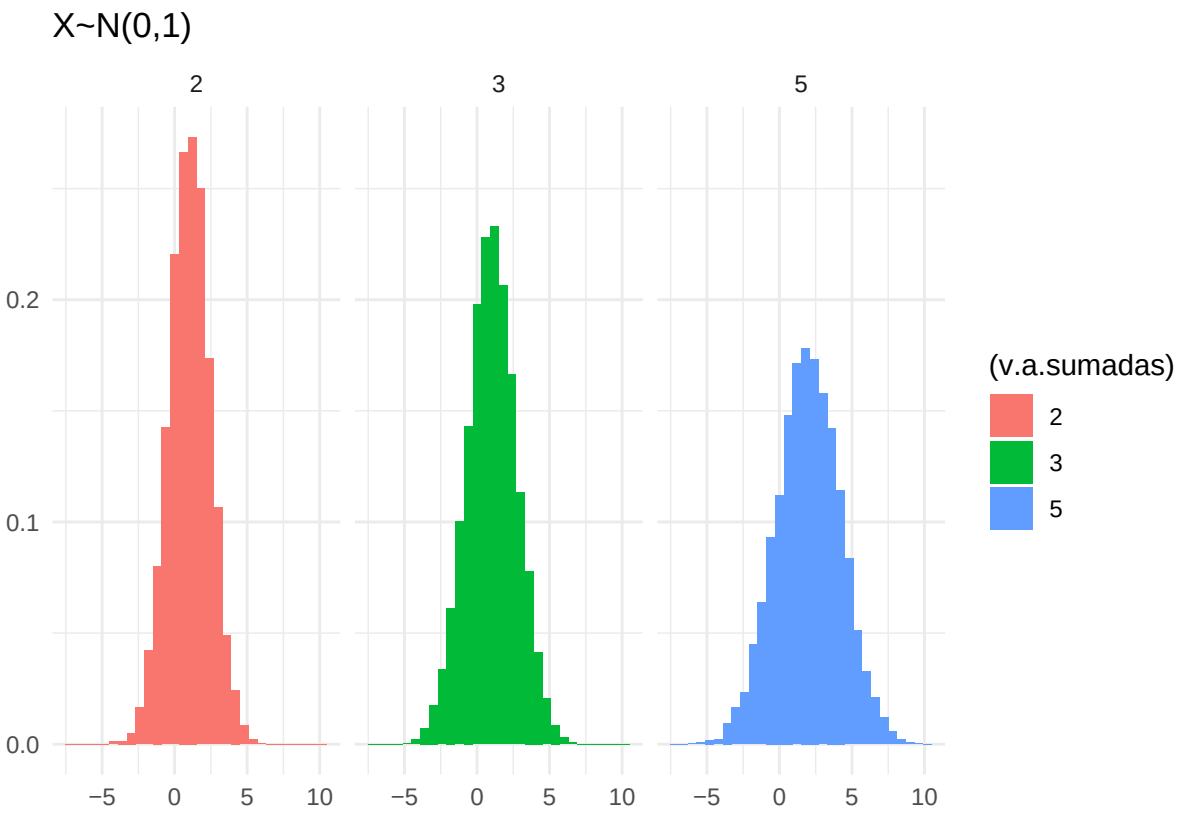
ii. $X_i \sim Exp(\lambda)$

```r
set.seed(123)
ensamble <- lapply(1:repeticiones,FUN=Simula,distribucion=rexp,
                   no.suma=c(2, 3, 5), parametros=c(1))
ensamble <- as.data.frame(t(as.data.frame(ensamble)))
ensamble$v.a.sumadas <- factor(ensamble$v.a.sumadas)
ggplot(ensamble, aes(x=realizacion, fill=(v.a.sumadas)))+
  geom_histogram(aes(y=..density..))+ facet_wrap(~v.a.sumadas, ncol=3) + theme_minimal() +ggtitle('X~Exp
```



iii. $X_i \sim N(0,1)$

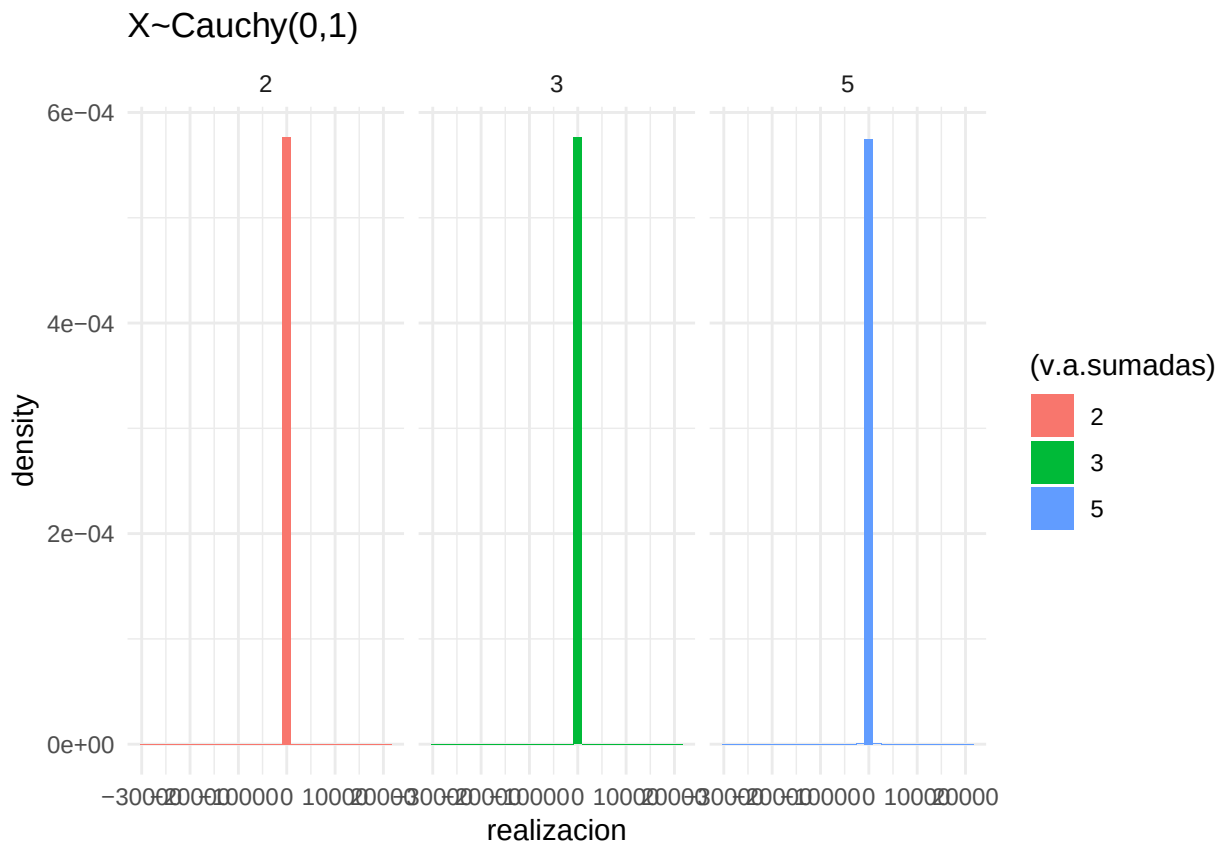
```r
set.seed(123)
ensamble <- lapply(1:repeticiones,FUN=Simula,distribucion=rnorm,
                   no.suma=c(2, 3, 5), parametros=c(0, 1))
ensamble <- as.data.frame(t(as.data.frame(ensamble)))
ensamble$v.a.sumadas <- factor(ensamble$v.a.sumadas)
ggplot(ensamble, aes(x=realizacion, fill=(v.a.sumadas)))+
  geom_histogram(aes(y=..density..))+ facet_wrap(~v.a.sumadas, ncol=3) +
  theme_minimal() +ggtitle('X~N(0,1)')+xlab('')+ylab('')
```



iv. $X_i \sim Cauchy(1)$

Como la distribución de Cauchy es estable la suma de i.i.d. se distribuye Cauchy, en la segunda gráfica se observa en escala logaritmica el eje x para apreciar mejor la distribución.

```
set.seed(123)
ensamble <- lapply(1:repeticiones,FUN=Simula,distribucion=rcauchy,
                   no.suma=c(2, 3, 5), parametros=c(0,1))
ensamble <- as.data.frame(t(as.data.frame(ensamble)))
ensamble$v.a.sumadas <- factor(ensamble$v.a.sumadas)
ggplot(ensamble, aes(x=realizacion, fill=(v.a.sumadas)))+
  geom_histogram(aes(y=..density..))+ facet_wrap(~v.a.sumadas, ncol=3) + theme_minimal() +ggtitle('X~Ca
```

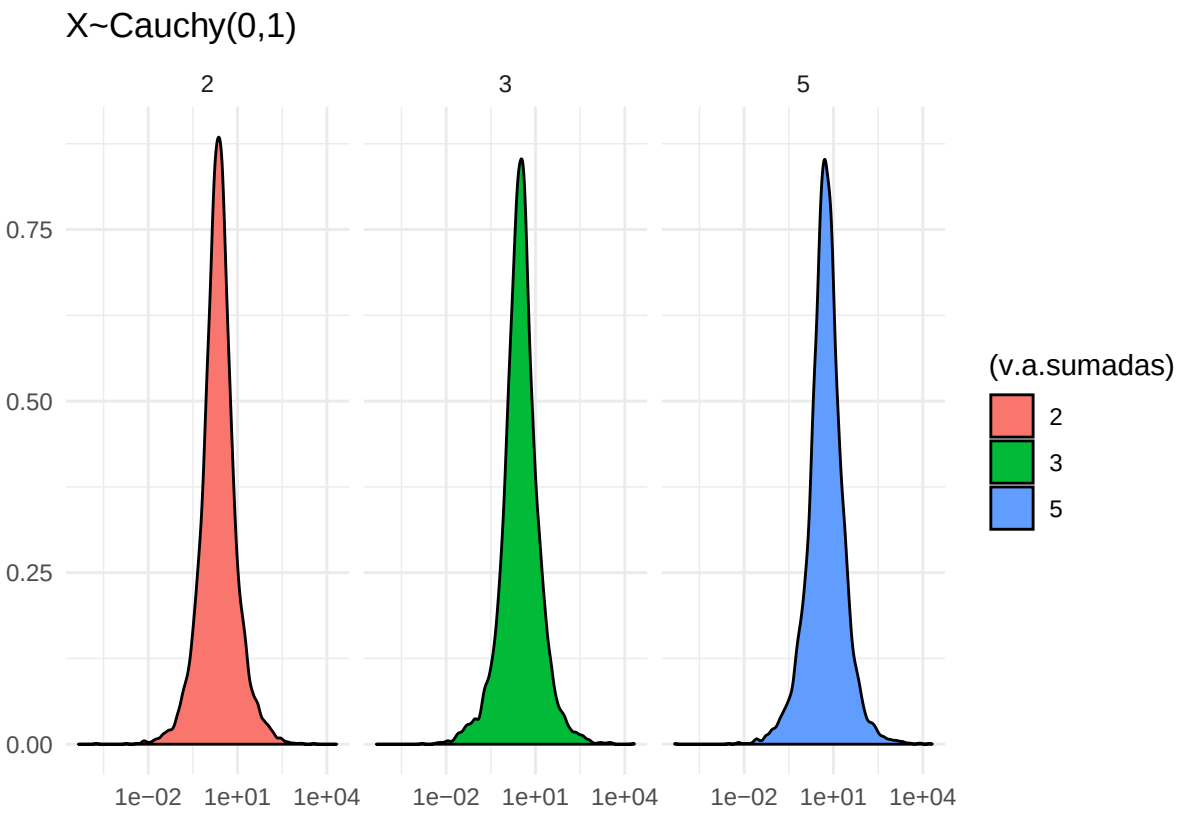



```
ggplot(ensamble, aes(x=realizacion, fill=(v.a.sumadas)))+
  facet_wrap(~v.a.sumadas, ncol=3) + theme_minimal() +
  ggtitle('X~Cauchy(0,1)')+geom_density() +
  scale_x_log10()+xlab('')+ylab('')
```

```
## Warning in self$trans$transform(x): NaNs produced
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 11084 rows containing non-finite values (stat_density).
```


X~Cauchy(0,1)
2
3
5
0.75
0.50
0.25
0.00
1e−02
1e+01
1e+04
(v.a.sumadas)
2
3
5

Ejercicio 3. Escribir un ensayo del artículo: Debashis Paul and Alexander Aue. Random matrix theory in statistics: A review (2014).

Sobre la importancia de las matrices aleatorias en el análisis de datos.

Una conciliación de mis dudas de estadística a lo largo de mi vida y un abrazo a esta nueva rama del conocimiento.

La estructura y estilo de este ensayo,ensayo porque postulo y defiendo una tesis, es particular debido a que en lugar de ser excesivamente técnico y enumerar comparaciones entre las pruebas "clásicas" de la estadística y el nuevo enfoque (y extensión) que presentan las matrices aleatorias presentadas en el paper de Debashis Paul and Alexander Aue, *Random matrix theory in statistics: A review* (2014) me concentro y narro algunas dudas referentes a matrices y estadística que he tenido desde los 24 años (cuando termine la licenciatura) hasta el día de hoy en mis estudios de posgrado y que las matrices aleatorias pueden responder, mostrando así que mi tesis, la cual es que **las matrices aleatorias generalizan resultados estadísticos y que su conocimiento es parte importante de mi desarrollo en la maestría** de cómputo de estadístico que estudio (no solo a nivel tecnico tambien personalmente me ayudara a conciliar conceptos) ya que muestran gran aplicabilidad.

La tesis la defiendo con argumentos que son anécdotas sobre diferentes dudas de matrices, me tomo la libertad de contextualizar el momento de las dudas, y dar una conclusión que el paper me recordó y en algunos casos aclaró.

Primero daré un contexto de mi background académico, por diferentes razones determine estudiar matemáticas "puras" en la facultad de ciencias, donde curse probabilidad y entonces me pareció una rama de las matemáticas más cercana al analisis matematico que a la geometría (en sus inicios me incliné fuertemente por las geometrías hasta que conocí las matemáticas discretas y aprendí a programar aunado al hecho de conocer al M. Lara Aparicio decidí cambiar de carrera ha matemáticas aplicadas). Como estudiante de la licenciatura en matemáticas aplicadas y computación me incliné hacia la estadística y las ciencias de la computación, tube la suerte de elegir las optativas de simulación, series de tiempo, análisis multivariado, estadística bayesiana y un montón de cursos de programación.

En lo que sigue presento mis experiencias (no en orden cronológico sino en el orden en que el paper me las originó cuando lo leí en el orden usual).

Al inicio de la sección 2 del paper me surge la primer experiencia **la distribución de los valores propios de una matriz**. Si bien en la maestría, en el curso de estadística multivariada, se requiere conocer los valores propios de diferentes matrices (de correlación, de varianza y otras) siempre los estimamos, el semestre pasado investigue un poco más sobre las pruebas para determinar la significancia de un valor propio en el contexto de PCA y la prueba que más me satisface consiste en simular matrices con la misma estructura que nuestros datos y checar la gráfica de los valores propios de la matriz simuladas contra los muestrales (que observamos en los datos); sin embargo esto solo permite descartar valores propios (de la misma manera que los test de esfericidad clásicos y que además asumen normalidad para sus resultados). En ese momento intuí que conocer la distribución de los valores propios de una matriz formalizará y resolvería el problema de la significancia de los vectores propios asociados al reducir dimensión en PCA, FA y aproximación de matrices. Si bien es un problema complejo este problema es atacado por esta nueva rama del conocimiento de la que estamos hablando, las matrices aleatorias. Sea de paso la duda sobre que tan util es un test de esfericidad sin los supuestos de normalidad también me surgió en el 2014 cuando realice mi servicio social en banxico y era importante descartar variables para poder analizar (por cuestiones computacionales) las mediciones en imágenes de billetes.

En la misma sección se habla sobre el modelo causal que se presupone en FA y que con una teoría de distribución de valores propios puede verse enriquecida pues los factores pueden resumirse en los valores

propios y tal vez esto de pie a reformulaciones del estudio estadístico que suele hacerse en psicología usando FA.

A finales de la segunda sección se habla de la distribución de Wishart, resultado importante en la base teórica de las matrices aleatorias, y recordé que esa fue la primer distribución que me causo un sentimiento de pánico pues trata sobre distribuciones de matrices. La distribución de una matriz es un tema que identifique en la licenciatura cuando curse análisis multivariado pues el supuesto del QDA es que las matrices de covarianza de las dos muestras nos diferentes lo cual me hizo investigar junto con mis amigos sobre una prueba formal para esta hipótesis la mejor prueba que encontramos fue diseñada por Levene sin embargo como resume la información de la matriz a su determinante no me dejo satisfecho en su momento, de nuevo las matrices aleatorias tienen herramientas más potentes y útiles, para resolver esto , este tema se retoma en el paper en la sección 4.1.1 con la ley de Tracy-Widom y en la sección 4.1.4 donde se presenta una corrección al test de máxima verosimilitud en su distribución χ^2 en dimensiones altas (justo el tema del test de máxima verosimilitud fue mi proyecto final de inferencia estadística y lo expuse formalmente y a pesar de ser el test uniformemente más potente para la familia exponencial pierde potencia en dimensiones altas).

Finalmente en la sección 3.1.4 se cita el problema de valores propios generalizado, un tema en el que me adentre en la materia de métodos numéricos con la finalidad de segmentar imágenes usando el algoritmo de normalized cut, en particular usando el método de Lanczos para encontrar los valores propios extremos (los más pequeños) que si bien no es propiamente estadístico es un problema de corte probabilístico pues la correctes y velocidad de convergencia de las variantes de los algoritmos de Lanczos y Arnoldi utilizan análisis matemático. Esto está intrínsecamente ligado, en mi experiencia, al problema de particionar grafos enormes (útil en el contexto de análisis de redes sociales), y para concluir este ensayo narro el contexto de cómo me encontré con este problema, en el 2017 un par de días antes de mudarme a la ciudad de Monterrey fui a recoger mi título universitario y mientras esperaba en la sala leí una tesis del CINVESTAV en CDMX que abordaba dos problemas el primero el de muestrear el grafo con el que se modela facebook y wikipedia y el segundo ¿como encontrar clusters de interés en ambas redes?, como el tramite del titulo en la UNAM es tardado termine de leer la tesis, cerré el archivo en mi celular, pensé en la cantidad de matrices que pueden formarse a partir de un número finito pero fijos de nodos donde las aristas tienen probabilidades de existir (no pensé en una distribución en la matriz de adyacencia de tales grafos) me alegré al pensar en los problemas que podría abordar después de concluir la MCE (y que las matrices aleatorias parecen ser de gran ayuda) y me serví un café en la sala de espera...