CIMAT

Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Distribución de Wishart y Marchenko-Pastur

Ricardo Cruz

31 de agosto de 2019

Al realizar el proceso de componentes principales, una de las tareas más importantes es determinar el número de componentes que se considerarán. Usualmente se optar por elegir esta cantidad por medio de un screeplot o tomar los primeros valores propios que acumulen el 80 o 90 % de la varianza.

Esto se puede ejemplificar con para la matriz de covarianzas gaussiana. Sea Z una matriz de $p \times n$, cuyas entradas se distribuyen normal estandar y D una matriz diagonal de dimensión $p \times p$. Sea $X = DZ$, entonces, XX' sigue una distribución Wishart con parámetros n y D^2. Sea $S = XX'/n$ la matriz de covarianza muestral, se obtienen los valores propios de esta matriz, los cuales contienen la varianza explicada por la respectiva componente.

Tomando $p = 30, n = \{30, 300\} D^2 = diag(12, 11, 10, 9, 8, 7, 3, 3, 3, \ldots, 3)$, los valores propios ordenados para $n = 30$ se muestran en la figura 1, mientras que la figura 2 los muestra para $n = 300$

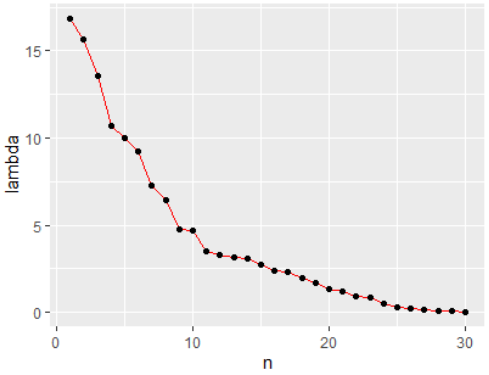

Figura 0.1: n=30.

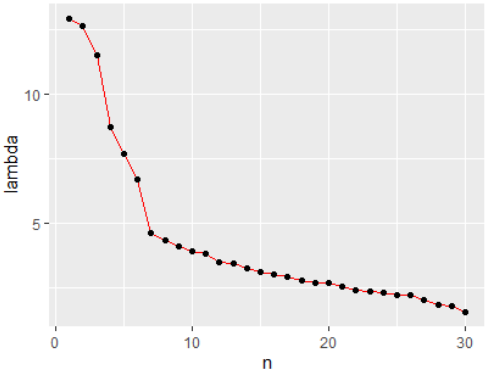

Figura 0.2: n=300.

Tomando el criterio del 80 %, se consideran 12 y 18 componentes para $n = 30$ y $n = 300$ respectivamente.
Para $n = 300$ simulamos vía Montecarlo los 30 valores propios de la matriz S. Considerando 1000 simulaciones, se tiene la densidad mostrada en la figura 3:

Cuando n, p son lo suficientemente grandes $n > p$, entonces $\frac{p}{n} \in (0, 1)$, entonces los valores propios de la matriz que sigue la distribución Wishart tienen una densidad espectral dada por:

$$\rho(\lambda) = \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{2 * \pi * c * \lambda}$$

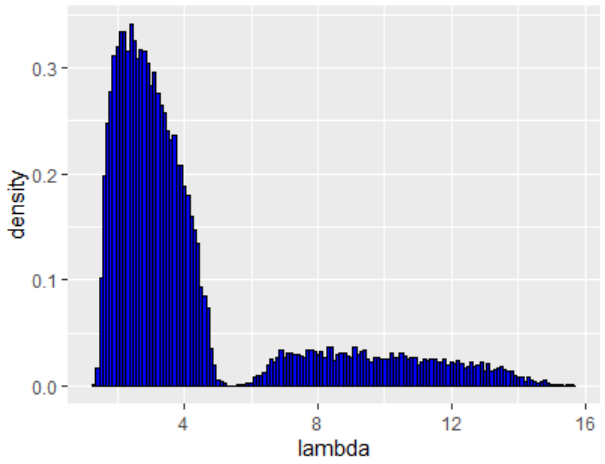

Figura 0.3: densidad de las simulaciones de los eigenvalores.

Donde $\lambda_{max} = (1 + \sqrt{c})^2$ y $\lambda_{min} = (1 - \sqrt{c})^2$. A esta densidad se le conoce como la ley de Marchenko-Pastur

Para el ejemplo que se ha manejado, $(n = 300)$ la figura 4 muestra la densidad vía la ley de Marchenko-Pastur y la figura 5 muestra las dos distribuciones hasta ahora generadas.

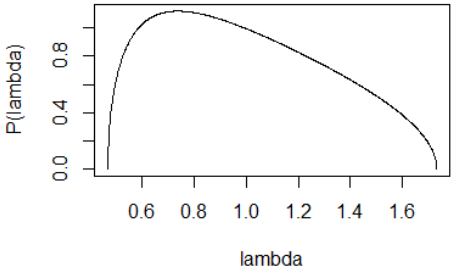

Figura 0.4: MarchenkO-Pastur.

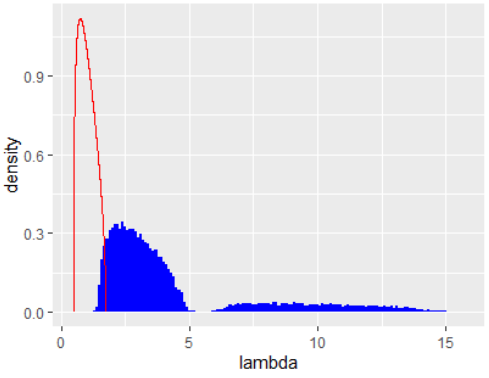

Figura 0.5: Distribuciones-

La ley de Marchenko-Pastur sugiere que los valores propios significativos (de los cuales se considerará la componente principal) son aquellos que son mayores al valor de λ_{max}.

Aplicando esta metodología, para $n = 30$ se tomarían 10 componentes, mientras que para $n = 300$ se toman 28 componentes. En ambos casos, se difiere

de lo planteado originalemente por el screeplot o el critério del 80 %

Ahora, para la matriz de correlación, las entradas que no se encuentran en la diagonal se distribuyen con media 0 y varianza $1/n$, por lo tanto, a medida que el tamaño de la muestra n aumenta, el valor de dicha entrada es 0 con mayor seguridad, puesto que la varianza se aproxima a 0 a medida que n crece y su media no se modificia.

Considerando una matriz aleatoria con $p = 40$ y $n = \{10, 1000\}$, las figuras 6 y 7 muestran los mapas de calor para las matrices de correlaciones con n igual a 10 y 1000 respectivamente.

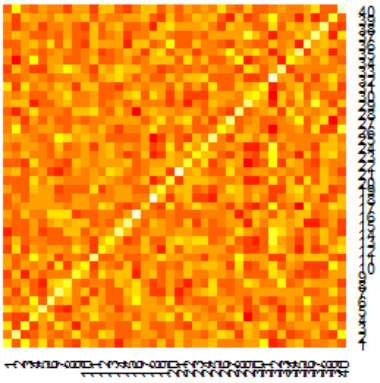

Figura 0.6: n=10

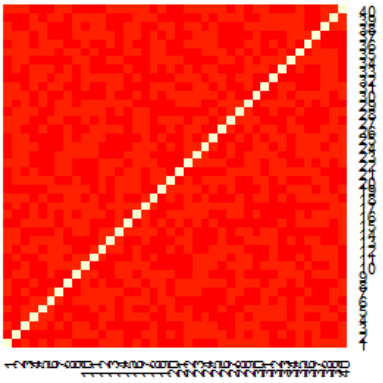

Figura 0.7: n=1000

Simulando 1000 veces la matriz de correlaciones, se puede obtener la media y desviación de cada entrada para los dos casos planteados para n.

Para $n = 10$ las entradas de la diagonal tienen una media al rededor de 1 y su desviación es aproximadamente .4. Para los elementos que no están en la diagonal, la media es aproximadamente 0.04 y desviación de .3

Para $n = 1000$ las entradas en la diagonal tienen media 1 y desviación aproximada de .04. Para los elementos en la diagonal, la media es aproximadamente 0, variando a partir de la cuarta posición después del punto y su desviación es aproximadamente .03.

Es por eso que el primer mapa de calor presenta mayor variación y el segundo

solo dos colores, 1 en la diagonal y 0 en el resto de las entradas. Este efecto se conoce como *noise dressing*, el cual depende del tamaño de la muestra.