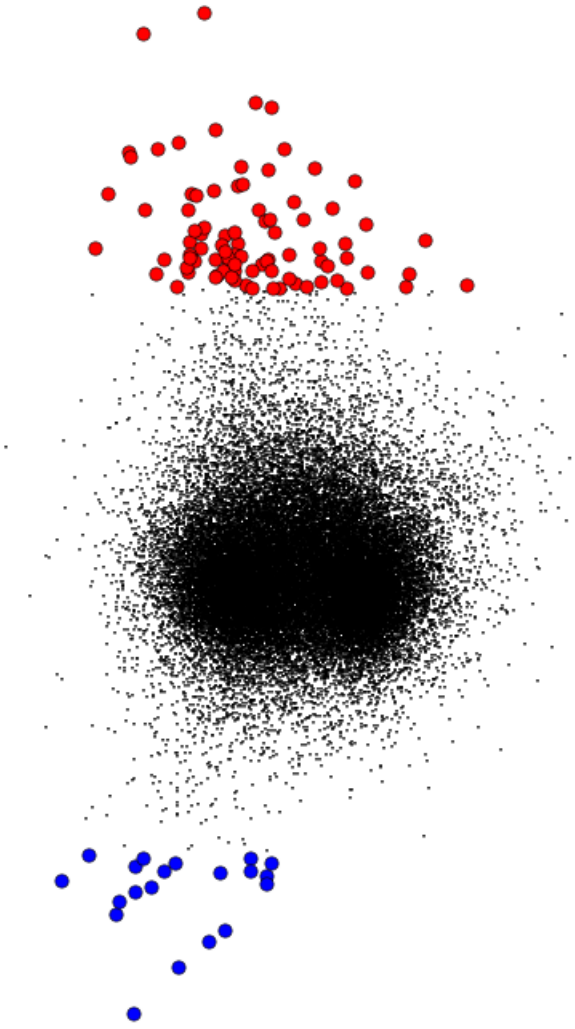
MÉTODO DE DETECCIÓN TEMPRANA DE OUTLIERS



Juan Gabriel Moreno Castellanos

Director de Trabajo de Grado: Milton Januario Rueda Varon Ph.D.

Codirectora de Trabajo de Grado: Luz Marina Moya Moya M.Sc.

Pontificia Universidad Javeriana

Mayo de 2012

.

Agradecimientos

Agradezco primero a Dios los resultados obtenidos y el logro que estos implican en mi vida académica. Agradezco también a mis padres por sus constante apoyo y por su soporte en cada momento, al Doctor Milton Rueda por el tema propuesto, por sus correcciones y constantes consejos, a la profesora Luz Marina Moya por su apoyo y sus correcciones, al Director del Departamento de Matemáticas, Fernando Novoa por brindar todas las herramientas para culminar este trabajo, al Director de Carrera Jesús Ochoa por su comprensión, a todos mis amigos por brindarme fuerza en sus palabras y a la Pontificia Universidad Javeriana por incentivar la investigación para el desarrollo de nuestro país.

.

Índice general

.

Introducción

Un outlier o dato extremo es una observación que se desvía de las otras o en otro sentido, datos que parecen inconsistentes con el conjunto de datos.

Los métodos para detección de outliers pueden ser clasificados como univariantes y multivariantes. En la práctica las variables tienen valores inusuales, muy grandes o muy pequeños; estos outliers pueden ser causados por medidas incorrectas, datos erróneos o por venir de una población diferente a la mayoría de datos. Los outliers causan efectos negativos en el análisis de datos, Osborne y Overbay (2004) [8] categorizan el perjudicial efecto de los outliers en los análisis estadísticos:

-Los outliers aumentan la varianza del error y reducen la potencia de las pruebas estadísticas.

-Si no son distribuidos aleatoriamente, pueden quebrantar la normalidad (en el análisis multivariado, violan los supuestos de esfericidad y normalidad multivariante). En pruebas de hipótesis las probabilidades pueden influenciar el error tipo I (rechazar una hipotesis que es verdadera) y tipo II (no rechazar una hipotesis que es falsa).

-Pueden alterar las estimaciones causando sesgos.

El siguiente ejemplo muestra como un valor extremo puede modificar las estimaciones puntuales de los parámetros, $\mu = E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$ y $\sigma^2 = E[(x-\mu)^2]$, y el intervalo de confianza del 95 % para μ.

Ejemplo 1 *Se supone que hay un conjunto de datos compuesto por* $1, 2, 3, 4, 5, 6, 7$. *En el cuadro 1 se observan las estimaciones enunciadas anteriormente:*

$E[X]$	Mediana	$V(X)$	Intervalo de confianza del 95 % para μ
4	4	4.67	[2,6]

Cuadro 1: Estimaciones en datos, sin Outlier.

Ahora si se reemplaza el punto 7 *por* 77*, en el cuadro 2.*

$E[X]$	Mediana	$V(X)$	Intervalo de confianza del 95 % para μ
14	4	774.67	[-11.74 , 39.74]

Cuadro 2: Estimaciones en datos, con outlier.

Como se puede observar en el Cuadro 2 el valor esperado y la varianza son modificados por un valor inusual, el 77. Además del intervalo de confianza se amplió debido al nuevo conjunto de datos. De esta manera, este dato podría causar problemas cuando el análisis de datos sea sensible al valor esperado o a la varianza. Estos outliers pueden brindar información útil acerca de los datos cuando se presentan respuestas inusuales en un estudio.

Ejemplo 2 *Los siguientes casos son situaciones en las cuales se encuentran outliers en términos de análisis específicos:*

-Los precios de ventas de un producto en la bolsa al finalizar una jornada.

-Identificar a los médicos que utilizan más o menos procedimientos específicos o equipos médicos, por ejemplo los instrumentos de rayos x.

-Identificar los profesores de determinada universidad con calificaciones altas de insatisfacción o un número de denuncias mayor al de otros profesores.

En la figura 1 los datos parecen estar distribuidos normalmente pero en ambos casos hay la presencia de outliers (circulos rojos).

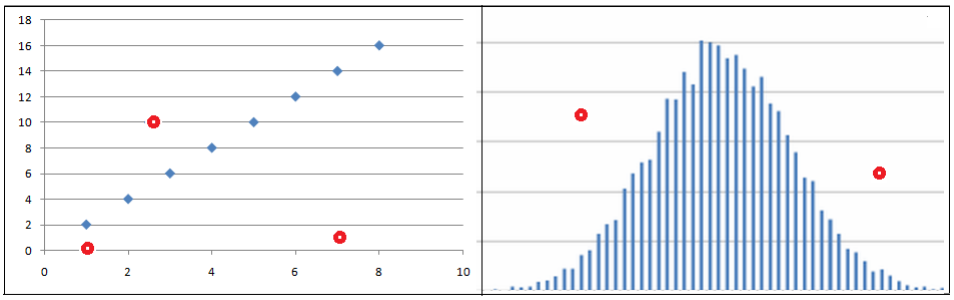


Figura 1: Outlier gráfico

En el presente trabajo, se realizará una revisión bibliográfica acerca de los métodos más utilizados para la detección de outliers. Además se propondrá un método dinámico univariante basado en indicadores robustos que faciliten dicha detección. Se diseñarán algunos estadísticos, luego se evaluará la pertinencia de estos y finalmente se analizará el comportamiento estadístico.

En el capítulo 1 se presentarán los métodos más destacados para detectar outliers, en la mayoría de estos se requiere información acerca de los datos y de la distribución de las variables. Además que sean i.i.d (independientes e idénticamente distribuidos), y muchas veces se asumen los parámetros y el tipo de outlier esperado. En otros métodos simplemente se requieren los datos sin ningún supuesto. En el capítulo 2 se nombran algunos de los métodos multivariantes más utilizados para la detección de outliers. En el capítulo 3 se describe la metodología propuesta. En el capítulo 4 se realiza una aplicación de esta metodología. Finalmente en el capítulo 5 se muestran los resultados y conclusiones.

Capítulo 1

Métodos univariantes de detección de outliers

1.1. Métodos basados en estadísticas

Este es un modelo que permite la generación de un pequeño número de observaciones, estas son muestras aleatorias de las distribuciones $G_1, G_2, ..., G_k$ las cuales difieren de la distribución objetivo F, que frecuentemente es tomada como $N(0,1)$. Este problema de identificación es trasladado a estudiar las observaciones que caen en la región de outlier, definida a continuación y propuesta por Davies y Gather (1993)[8].

Definicion 3 *$\forall\alpha$ llamado coeficiente de confianza tal que $0 < \alpha < 1$, la región de outlier de $N(0,1)$ es:*

$$out(\alpha, \mu, \sigma_2) = \{x : |x - \mu| > z_{1-\alpha/2}\sigma\}$$

Donde Z_q es el q quintil de la $N(0,1)$.

Se concluye que x es un outlier respecto a F si $x \in out(\alpha, \mu, \sigma^2)$. Esta técnica es basada en la construcción del intervalo de confianza, véase la figura 1.1.

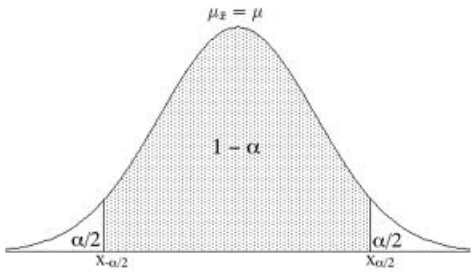


Figura 1.1: Región de outlier.

1.2. Métodos paso simple y secuenciales

1.2.1. Métodos paso simple

Estos métodos identifican outliers a la vez que sucesivamente eliminan o adicionan datos con respecto a la ecuación anterior, es decir determinando el α-región de outlier. Un identificador de paso simple esta dado por:

$$out(\alpha_n, \hat{\mu_n}, \hat{\sigma_n}) = \{x : |x - \alpha| > g(n, \alpha_n)(\hat{\sigma_n})\}$$

En este caso n es el tamaño de la realización (muestra), $\hat{\mu_n}$ y $\hat{\sigma_n}$ son los parámetros de la distribución objetiva basada en la muestra, α_n es el coeficiente de confianza y $g(n, \alpha_n)$ son los limites (números críticos de la desviación estándar) de la región de outlier.

Dado que $\hat{\mu_n}$ y $\hat{\sigma_n}$ son afectados por la presencia de outliers, el método podría generar falsos positivos, para évitar este problema, el α-valor debe disminuirse para tener en cuenta el número de comparaciones realizadas. Hay una aproximación denominada corrección de Bonferroni la cual fija el α-valor para el conjunto de n comparaciones igual a α, al tomar el α-valor para cada comparación igual α/n.

Además de esto el valor $g(n, \alpha_n)$ es especificado por métodos numéricos, tal como la simulación Monte Carlo (el modelo de Monte Carlo es un método no determinístico o estadístico numérico usado para aproximar expresiones matemáticas complejas y difíciles de evaluar con exactitud) para diferentes tamaños de la muestra.

1.2.2. Métodos de procesos secuenciales

Estos procesos se realizan paso a paso, y cada observación es examinada para determinar si es outlier. Estos procesos secuenciales se pueden clasificar como procesos interiores y procesos exteriores, véase la figura 1.2:

Proceso interior

En el proceso interior o también nombrado de selección progresivo, se toma a cada paso la más extrema de las observaciones, es decir, la que tiene la mayor medida atípica para luego ser examinada.

Proceso exterior

En el proceso externo la muestra de observación es reducida a una muestra más pequeña, por ejemplo un factor de dos, las observaciones removidas son guardadas en un depósito, las estadísticas se calculan en la muestra reducida y se remueve la observación del depósito, dicha observación es examinada en orden opuesto para indicar si corresponden a outliers.

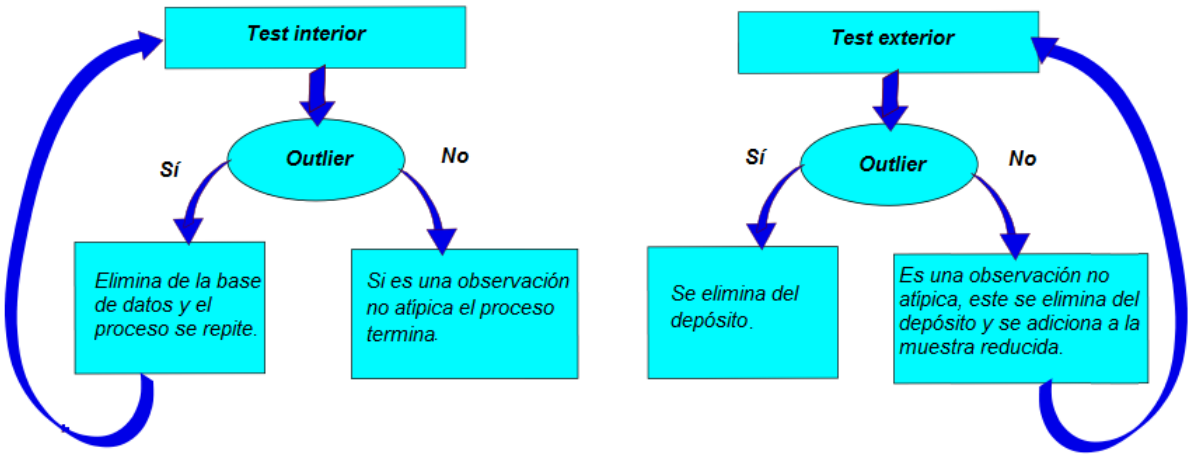


Figura 1.2: Proceso interno y Proceso externo.

1.3. Medidas robustas univariantes

Las estadísticas robustas proporcionan un enfoque alternativo a los métodos estadísticos clásicos. La motivación es la de producir estimadores que no se vean afectados indebidamente por desviaciones de los supuestos del modelo. Es claro que los parámetros μ y σ son eficientes si no hay presencia de outliers, es decir si la base de datos está contaminada, estos parámetros se pueden desviar de la detección deseada para outliers.

De acuerdo a los problemas de influencia de los outliers sobre los parámetros, Hampel (1971-1974) [6] define el concepto de *punto de quiebre* (breakpoint). Este punto es una medida de robustez de los estimadores contra outliers, se define como el menor porcentaje de outliers que pueden causar que un estimador tome valores arbitrariamente grandes. En consecuencia con esto, Hampel sugiere hablar de la desviación mediana absoluta (MAD).

1.3.1. MAD (Desviación Mediana Absoluta)

En estadística la desviación mediana absoluta es una medida robusta de la variabilidad de una muestra univariante de datos cuantitativos. También puede hacer referencia al parámetro de la población que se estima por el MAD calculada a partir de una muestra. Formalmente para un conjunto de datos univariados $X_1, X_2, ..., X_n$, el MAD se define como la mediana de las desviaciones absolutas con respecto a la mediana de los datos.

$$MAD = mediana_i(|X_i - mediana_j(X_j)|)$$

Es decir, a partir de los residuos (desviaciones) de la mediana de los datos, el MAD es la mediana de los valores absolutos. Este es un estimador robusto de localización y de propagación. Este método en gran medida no es afectado por la presencia de outliers en el conjunto de datos

aunque haya datos sesgados. Lo interesante del método es que utiliza la mediana en vez de la media y la desviación estándar.

Ejemplo 4 *Se consideran los datos* $1, 1, 2, 2, 4, 6, 9$.

$$mediana_j(x_j) = 2$$

Tiene una mediana de 2.
Ahora se calculan Las desviaciones absolutas i.e.

$$|x_i - 2|$$

y se obtiene $1, 1, 0, 0, 2, 4, 7$

Finalmente se calcula

$$MAD = mediana(1, 1, 0, 0, 2, 4, 7) = 1$$

Así que la desviación mediana absoluta de estos datos es 1.

Se introduce ahora un outlier igual a 23 en vez del 9. El resultado para MAD es el mismo. La primer mediana da lo mismo que la anterior

$$mediana_j(x_j) = 2$$

Ahora se calcula Las desviaciones absolutas i.e.

$$|x_i - 2|$$

$(1, 1, 0, 0, 2, 4, 21)$

Finalmente se calcula

$$MAD = mediana(1, 1, 0, 0, 2, 4, 21) = 1$$

Luego se puede observar que este operador no se ve afectado por outliers.

1.4. Diagrama de caja (Boxplot)

Un diagrama de caja es un grafico basado en cuartiles, propuesto por (Turkey(1977)) [24], mediante el cual se visualiza un conjunto de datos. Está compuesto por un rectángulo, la caja, dos brazos y los bigotes. Este grafico suministra información sobre los valores máximos y los mínimos, los cuartiles Q_1, Q_2 o mediana y Q_3, además de la existencia de outliers y la simetría de la distribución. El primer y el tercer cuadrante Q_1, Q_3 son usados para obtener las medias robustas para el valor esperado.

$$\hat{\mu_n} = (Q_1 + Q_3)/2$$

$$\hat{\sigma_n} = Q_3 - Q_1$$

Se habla de dos tipos de outliers los cuales pueden ser distinguidos, los outliers leves y los outliers *extremos*. Una observación es declarada como outlier extremo si esta cae fuera del intervalo $(Q_1 - 3IQR, Q_3 + 3IQR)$. Donde $IQR = Q_3 - Q_1$ llamado radio intecuartilico. O así mismo una observación x es declarada como un outlier *leve* si cae fuera del intervalo $(Q_1 - 1,5IQR, Q_3 + 1,5IQR)$, donde 3 y 1,5 son escogidos por comparaciones con una distribución normal.

Gráficamente la observación que esté fuera de las *vallas* superior o inferior es tomado como un outlier en potencia, véase la figura 1.3. Si no se garantiza la normalidad no influye en los resultados pues este depende de la mediana y no del valor esperado de los datos. Es extensamente usado en la detección de outliers, quizá el más utilizado y popular entre todos los métodos.

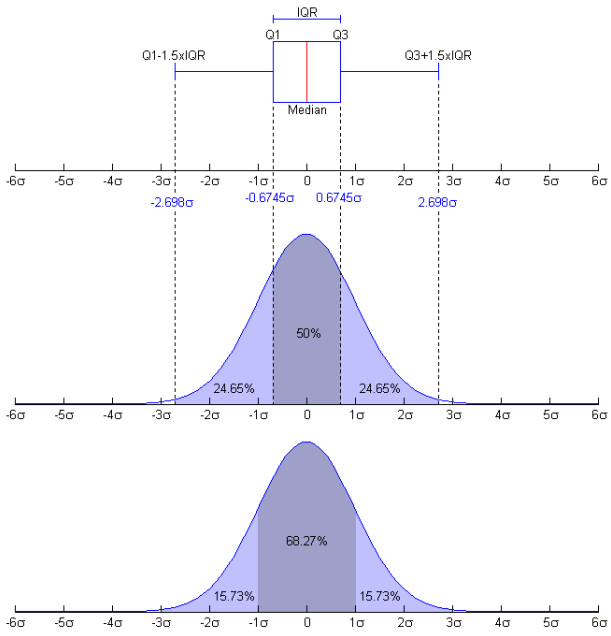


Figura 1.3: Diagrama de caja (Boxplot).

Ejemplo 5 *Se suponen unos datos de manera que*

$q_1 = 7$

$q_2 = 8{,}5$

$q_3 = 9$

Rango intercuartilico (RIC): $(q_3 - q_1) = 2$.

Para dibujar los bigotes, las líneas que se extienden desde la caja, se debe calcular los límites superior e inferior, LI y LS, los cuales identifican a los outliers. Se considera outliers, aquellos inferiores a $q_1 - 1,5 \cdot RIC$ *o superiores a* $q_3 + 1,5 \cdot RIC$.

se obtiene:

$LI = 7 - 1,5 \cdot 2 = 4$

$LS = 9 + 1,5 \cdot 2 = 12$.

Ahora se buscan los últimos valores que NO son atípicos, que serán los extremos de los bigotes. En este caso 5 y 10. A continuación se marcan los datos que están fuera del intervalo (LI, LS)*, es decir, los datos 0,5 y 3,5. Además, se pueden considerar valores extremadamente atípicos aquellos que exceden* $q_1 - 3 \cdot RIC$ *o* $q_3 + 3 \cdot RIC$. *En este caso:*

inferior $7 - 3 \cdot 2 = 1$

superior $9 + 3 \cdot 2 = 15$.

El valor 0,5 sería atípico extremo y el 3,5 seria atípico, es decir, estos son los outliers. En la figura 1.4 se muestran los outliers encontrados.

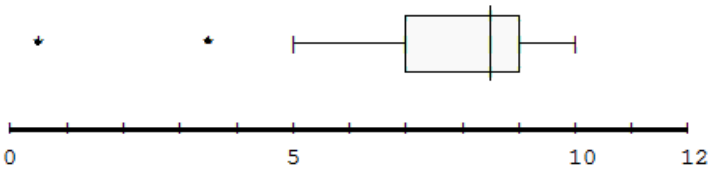


Figura 1.4: Boxplot Ejemplo 3.

1.5. Filtro-Outlier resistente en linea

Liu en 2004 [14] propuso un filtro-outlier robusto basado en el trabajo de Martin y Thomson en 1982 [16]. La propuesta del filtro limpiador incluía un resistente de outliers en línea, un estimador de los procesos de modelos que combinaba este con un filtro de Kalman (el filtro de Kalman es un algoritmo desarrollado por Rudolf Kalman en 1960 [8] que sirve para identificar el estado oculto, no medible, de un sistema dinámico lineal) modificado para detectar y *limpiar* outliers,

no sobra decir que no utiliza información a priori del modelo. Además detecta y reemplaza outliers en línea mientras reserva la otra información de los datos. Este método es eficiente para la detección de outliers y limpieza de datos para auto correlación e incluso procesos de datos no estacionarios.

1.6. SPC (Control de Procesos Estadísticos)

El proceso de control estadístico incluye herramientas como las cartas de control. El control estadístico de procesos (SPC) es la aplicación de métodos para el seguimiento y control de un proceso que asegura su máximo potencial para producir productos conformes.

Cartas de Control

Las cartas de control son una herramienta poderosa para analizar la variación de los datos. Estas enfocan la atención hacia las causas especiales de variación y reflejan de la variación debida a las causas comunes. Se dice que un conjunto de datos está bajo control estadístico cuando presenta únicamente causas comunes. En este caso tenemos una base de datos estable y predecible. Cuando existen causas especiales en la base de datos, esta fuera de control estadístico; los gráficos de control detectan la existencia de estas causas en el momento en que se dan, lo cual permite que podamos tomar acciones a tiempo. En análisis de cartas de control, es relevante el estudio de la variación en la información. Hay varios métodos de cartas de control, independientes, basados en variables y en atributos.

Se destacan los métodos siguientes:

- Carta CUSUM(Cumulative sum): Suma acumulada. Realiza un proceso de relaciones independientes, el tipo de proceso de observación es de atributos o variables.
- Carta EWMA(Exponentially weighted moving average): Medias móviles con ponderación exponencial. Este realiza un proceso de relaciones independientes, el tipo de proceso de observación es de atributos o variables.
- Carta de SERIES DE TIEMPO: Realiza un proceso de relaciones de autocorrelación, el tipo de proceso de observación es de atributos y variables.
- Carta de REGRESION: Realiza un proceso de relaciones dependiente de las variables de control del proceso, el tipo de proceso de observación es de variables.

1.7. Métodos tradicionales SPC

1.7.1. CUSUM (Cumulative Sum)

La suma acumulada de cartas de control propuesta por Shewart (1954) [9],[22], supone el cálculo de una suma acumulada (secuencial). A las muestras de un proceso de X_n se les asignan pesos ω_n y es resumido de la siguiente manera:

$$S_0 = 0,$$

$$S_{n+1} = max\{0, S_n + X_n - \omega_n\}.$$

Cuando el valor de S supera el umbral, un cambio en el valor se ha encontrado. La formula anterior solo detecta los cambios en la dirreccion positiva. Cuando los cambios son negativos hay que encontrar el mínimo en lugar del máximo y esta vez un cambio se ha encontrado cuando el valor de S es inferior al valor (negativo) del umbral. Además CUSUM no requiere de la función de verosimilitud.

Las sumas acumuladas toman una suma de las diferencias de cada variable aleatoria con el $E[X]$ de la variable de interés X hasta un índice, por ejemplo n, i.e $S_n = \sum \bar{X}_i - E[X]$. Si dicho proceso está bajo control, $E[X]$ es μ_0, así al hallar el valor esperado de el último término de la suma, S_n, se obtiene el valor cero y la interpretación es que se desarrolla alrededor de cero. Se puede decir que el proceso está bajo control. El proceso evoluciona de forma aleatoria alrededor de una horizontal a nivel cero. En este caso diríamos que no hay presencia de outliers, como puede verse en la figura 1.5.

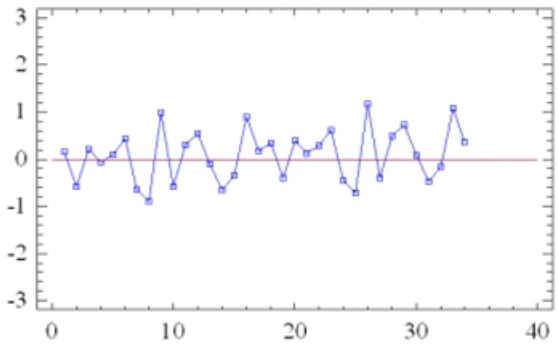


Figura 1.5: Cusum.

Si por el contrario el proceso no está bajo control, es decir, si se desajusta la medida del valor esperado de la v.a. de interés a $\mu_0 + k$, entonces el valor esperado del último término de nuestra suma acumulada (de nuevo se llama S_n) será igual a nk, donde k tiene una tendencia lineal, creciente si $k > 0$ o decreciente si $k < 0$. El proceso se desajusta, S_n evoluciona de forma aleatoria alrededor de una pendiente, como se ve en la figura 1.6.

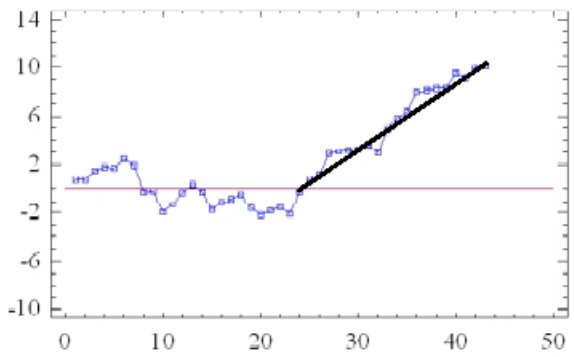


Figura 1.6: Cusum con desajuste.

La idea general es dar una alerta mediante un gráfico CUSUM en el cual, se traza una plantilla en V para tomar los límites y determinar el momento del desajuste o de la aparición de alguna observación atípica, es decir, la presencia de un outlier. Ver en la figura 1.7.

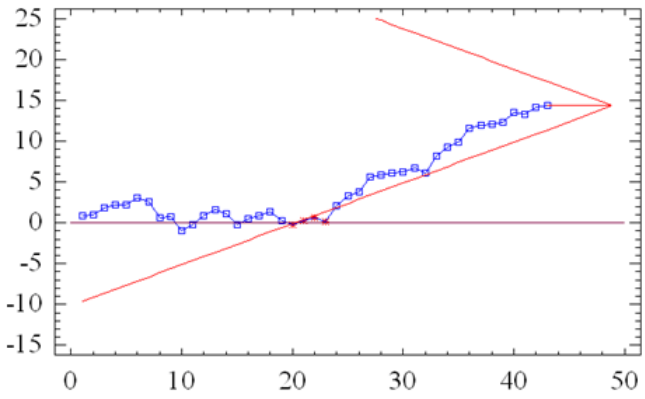


Figura 1.7: Detección del outlier.

Al realizar un gráfico que muestra las desviaciones positivas y negativas, se puede ver la alerta y el momento del desajuste, véase la figura 1.8

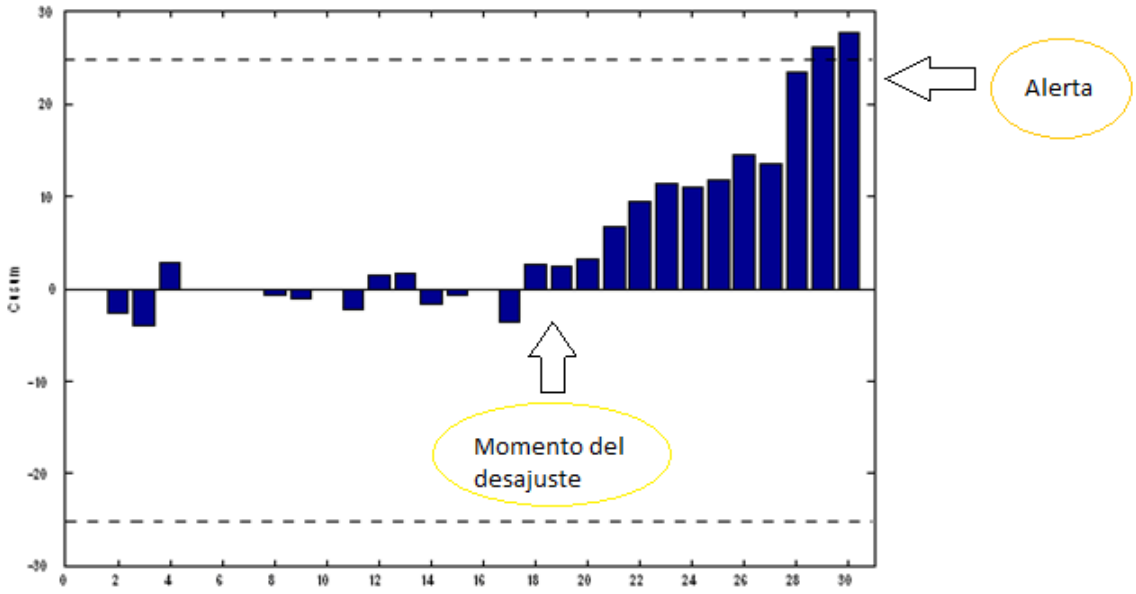


Figura 1.8: Alerta del momento del desajuste

1.7.2. EWMA (Medias Móviles con Ponderación Exponencial)

Sea X la variable de calidad de interés, el gráfico EWMA representa la evolución del estadístico.

$$y_i = \lambda x_i + (1 - \lambda) y_{i-1}$$

Con $y_0 = \mu_0$ (esto lo decide el analista) y $0 < \lambda \leq 1$. La mayoría de los métodos para datos dependientes son basados en series de tiempo.

EWMA es un gráfico de control utilizado para monitorear variables o atributos de bases de datos, podría ser la historia de todo un proceso industrial de producción. Otros gráficos de control tratan subgrupos racionales de forma individual, el gráfico EWMA rastrea la media móvil ponderada exponencialmente de todos los medios de la muestra anterior.

Los pesos EWMA de la muestra se ubican geométricamente en orden decreciente de forma que las muestras más recientes se ponderan más altas, mientras que las muestras más distantes contribuyen muy poco (estos pesos decaen exponencialmente). Cuando se le da un valor de peso al parámetro λ se tienen dos situaciones: λ es cercano a 1, en este caso la memoria del proceso es corta, el pasado tiene poco valor, pero si al contrario el valor λ es cercano a cero, los pesos decaen más despacio, el pasado tiene valor, la memoria es larga.

La figura 1.9 muestra como decaen los pesos exponencialmente y el efecto de la selección de λ:

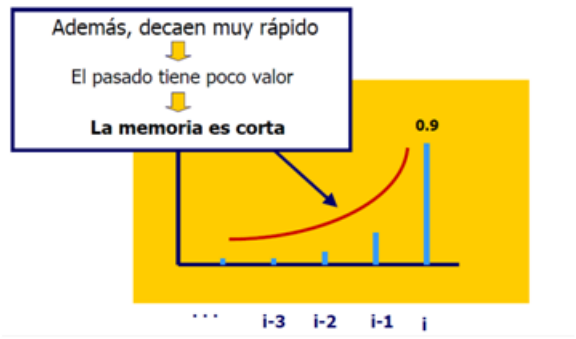


Figura 1.9: EWMA con λ cercano a 1.

Si λ es cercano a 1 (en la figura 1.9 para $\lambda = 0,9$) hay poca memoria.

Figura 1.10: EWMA con λ cercano a 0.

Si λ es cercano a cero (para la figura 1.10 con $\lambda = 0,1$) hay memoria. Así para detectar un desajuste pequeño, se necesita acumulación de información de varios periodos para que el desajuste pueda apreciarse. Si queremos detectar desajustes grandes, bastará con la información inmediatamente posterior al desajuste, no se debe acumular información. Mediante algunos cálculos adicionales es posible graficar algunas fronteras (límites superior, inferior y central), véase la figura 1.11.

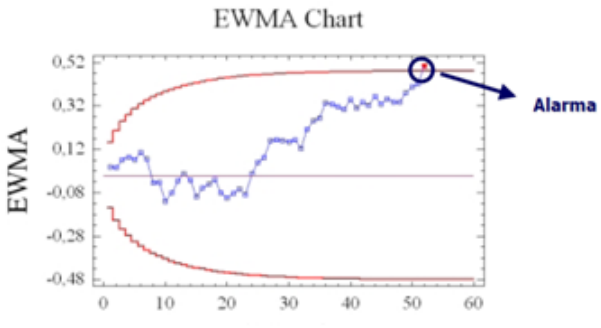


Figura 1.11: Deteccion del EWMA.

Este método indica mediante una alarma cuando hay algo mal en los datos, es decir la presencia de algún outlier, esto nos lleva a su detección.

1.8. Modelos ARIMA (Modelo de Autoregresión de Promedios Moviles Integrado)

Es un modelo para datos dependientes. Una familia de modelos implementados para la estimación y filtración de procesos de autorelación, bajo ciertas suposiciones, los residuos del modelo ARIMA son independientes y distribuidos aproximadamente normal, al cual se le puede aplicar el tradicional SPC. De los métodos ARIMA podemos destacar el AR (autoregresivo) y el MA (promedio móvil) para la detección de outliers.

1.8.1. AR (Autoregresivo)

Es un modelo aleatorio que se utiliza para modelar y predecir distintos tipos de fenómenos naturales. El método AR es una forma alternativa para calcular el espectro de señales. Es especialmente útil cuando las señales tienen una baja relación señal/ruido. Formalmente:

$AR(p)$ se refiere al modelo de autoregresión de orden p, se escribe:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

donde $\varphi_1, \varphi_2, ..., \varphi_p$ son parámetros del modelo, c es una constante y ε_t es ruido blanco.

Algunas limitaciones en los valores de los parámetros son necesarias para que el modelo se mantenga estacionario. En este caso el modelo predice la aparición de outliers en la base de datos.

1.8.2. MA (Promedio Movil)

En el análisis de series de tiempo, el promedio móvil (MA) es un enfoque común para la modelización univariante de series de tiempo. La notación $MA(q)$ se refiere al modelo de promedio móvil de orden q:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}$$

donde μ es la media de la serie, $\theta_1, ..., \theta_q$ son los parámetros del modelo y $\varepsilon_t, \varepsilon_{t-1}, ...$ son ruido. El valor de q se llama el orden del modelo MA. Un modelo de media móvil es una regresión lineal del valor actual de la serie contra el anterior (no observada) en términos de ruido blanco o perturbaciones aleatorias. Los choques al azar en cada punto, se supone, se producen con la misma distribución, por lo general una distribución normal, con ubicación en la escala de cero y una constante. En este modelo los choques aleatorios se propagan a los valores futuros de la serie de tiempo.

Capítulo 2

Métodos multivariantes de detección de outliers

Comunmente las observaciones multivariantes no pueden ser detectados en cada variable, o en simples univariantes independientes. La detección de outliers en multivariantes solo se puede desarrollar si ya se ha hecho una relación entre las diferentes variables. Un ejemplo de una mala detección de un outlier en multivariantes es el que se muestra en la figura 2.1

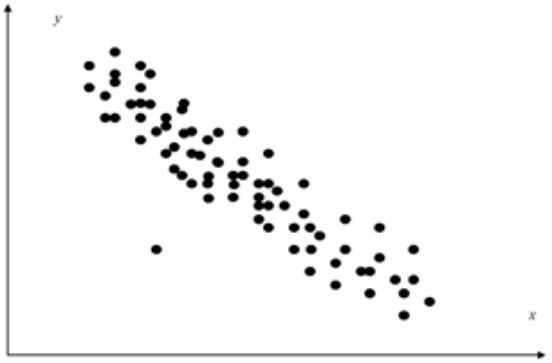


Figura 2.1: Outlier multivariante invisible al estudio de univariantes

Si se estudian estos datos como univariantes independientes jamás se encontrará el dato anormal que se ve en la parte baja izquierda de la figura. Además de esto se presentan dos efectos cuando existen múltiples outliers en los casos multivariados:

Efecto de enmascaramiento

Cuando un outlier esconde un segundo se dice que hay un efecto de enmascaramiento, el segundo outlier se puede ver como outlier si se encuentra solo, si está en compañía del primer outlier este se toma como una observación no atípica. Luego de que se elimine el primer outlier, el segundo será visto. Este enmascaramiento se produce cuando un grupo de observaciones atípicas sesga la media y la covarianza estimada, y la distancia que resulta del punto periférico de la media es pequeña.

Efecto de inundamiento

Cuando un outlier inunda una segunda observación se dice que hay un efecto de inundamiento, es decir, el segundo outlier es considerado como tal si está acompañado del primer outlier. Si se elimina el primero, el segundo se convertirá en una observación no atípica. Esto ocurre cuando un grupo de instancias periféricas sesga la media y la covarianza estimada y lejos de otros casos no periféricas, la distancia resultante de estos casos a la media es grande, haciendo que se vean como los valores extremos.

Se pueden destacar los siguientes métodos de detección multivariante:

Métodos estadísticos:

- Detección de outliers basado en estadística robusta.

Métodos de ´Data Mining´:

- Detección de outliers por clustering (agrupamiento).
- Detección de outliers basado en distancia.
- Detección local de outliers basado en densidad.

2.1. Métodos estadísticos:

2.1.1. Detección de outliers basado en estadística robusta

Esta detección da una aproximación alternativa a los métodos estadísticos clásicos. La motivación es la producción de estimadores que no se vean afectadas indebidamente por pequeñas desviaciones de los supuestos del modelo, como por ejemplo los outliers.

Las estadísticas robustas buscan proveer métodos que emulen métodos de estadística populares, pero que no sean indebidamente afectados por outliers u otras pequeñas salidas de la suposición del modelo. Se suele suponer que los residuos de los datos son normalmente distribuidos.En caso contrario aplicando el teorema central del límite y producir estimaciones de una distribución normal. Cuando hay presencia de outliers en los datos, los métodos clásicos son pobres en términos de rendimiento, como por ejemplo el filtro de Kalman que no es robusto.

El uso de estimaciones sólidas de los parámetros de distribución multidimensional puede mejorar el rendimiento de los procedimientos de detección de presencia de outliers. Hadi (1992) [8]aborda este problema y propone que se sustituya el vector de medias por un vector de variables medianas y calcular la matriz de covarianza para el subconjunto de estas observaciones con la menor distancia de Mahalanobis [15].

En estadística, la distancia de Mahalanobis es una medida que dermina la similitud entre dos variables aleatorias multidimensionales. Se diferencia de la distancia Euclídea en que tiene en cuenta la correlación entre las variables aleatorias.

Definicion 6 *La distancia de Mahalanobis entre dos variables aleatorias con la misma distribución de probabilidad para $\vec{x}$ y $\vec{y}$, con matriz de covarianza $\sum$ se define como:*

$$d_m(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T (\sum)^{-1} (\vec{x} - \vec{y})}$$

Tal como una métrica.

Caussinus y Roiz (1990) [2], [8] proponen una estimación sólida de la matriz de covarianza, que se basa en observaciones ponderadas de acuerdo a su posición del centro. Los autores también proponen un método de proyecciones de baja dimensión del conjunto de datos. Usan el principio generalizado de análisis de componentes (GPCA) para revelar las dimensiones que muestran los outliers. Otros estimadores robustos de la ubicación (centroide) y la forma (matriz de covarianza) son el factor determinante de covarianza mínimo (MCD) y el elipsoide de volumen mínimo (MVE) propuesto por Rousseeuw (1985)[21], Rousseeuw y Leory (1987) [8], y Acuña y Rodríguez (2004) [1].

MVE (Minimum Volume Ellipsoid) El Valor Mínimo del Volúmen del Elipsoide

Es un estimador que minimiza el volumen de la matriz de covarianza asociada a la submuestra, se basa en el elipsoide de menor volumen que cubre h de las n observaciones. Se trata de un equivariante afín, un estimador robusto de alto desglose de la ubicación de múltiples variables y de dispersión. El MVE puede calcularse mediante un algoritmo de remuestreo. Su sesgo de baja hace que sea útil para la detección de outliers en los datos multivariados, mediante el uso de las distancias robustas basadas en MVE.

Ejemplo 7 *Los puntos suspensivos representan el 97,5 % de tolerancia del elipsoide en las observaciones, sobre la base de la media muestral y la matriz de covarianza de la muestra, como se ve en la figura 2.2.*

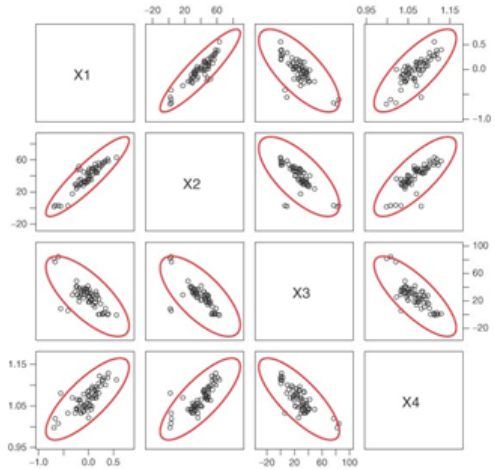


Figura 2.2: Ejemplo MVE.

MCD (Minimum Covariance Determinant) El determinante de covarianza mínima

El estimador determinante de covarianza mínima (MCD) es un estimador multivariable robusto, de ubicación y dispersión. Puede calcularse de manera eficiente con el algoritmo de FAST-MCD de Rousseeuw y Van Driessen [25]. Dado que la estimación de la matriz de covarianza es la piedra angular de métodos estadísticos multivariados, el MCD se ha utilizado también para desarrollar técnicas multivariantes robustas y eficientes computacionalmente, véase la figura 2.3.

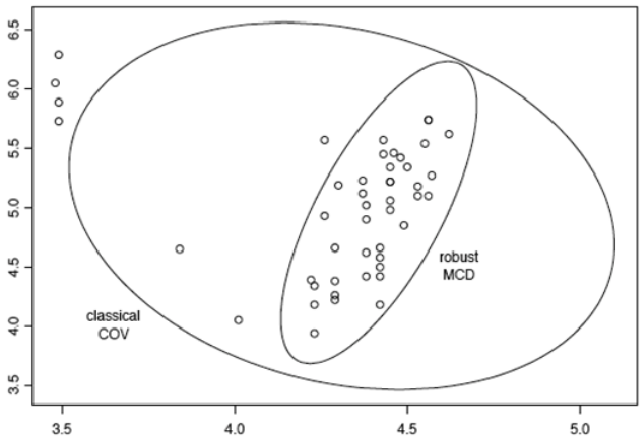


Figura 2.3: MCD.

2.2. Métodos de data mining.

En contraste con los métodos estadísticos, los métodos relacionados con minería de datos son no paramétricos, asume ninguna información del modelo, estos métodos son perfectos para conjuntos de datos de altas dimensiones, de estos metodos puede reconocer categorías que se estudiaran a continuación:

2.2.1. Métodos basados en clustering

Son utilizados como una herramienta independiente para obtener conocimiento sobre la distribución de un conjunto de datos, por ejemplo, enfocar su análisis y procesamiento de datos, o como un paso del proceso previo de otros algoritmos que operan en los grupos detectados.

Hay diversas técnicas, algunas tienen en cuenta distancias, densidades y otros ubicaciones en regiones especificas. Vease la figura 2.4.

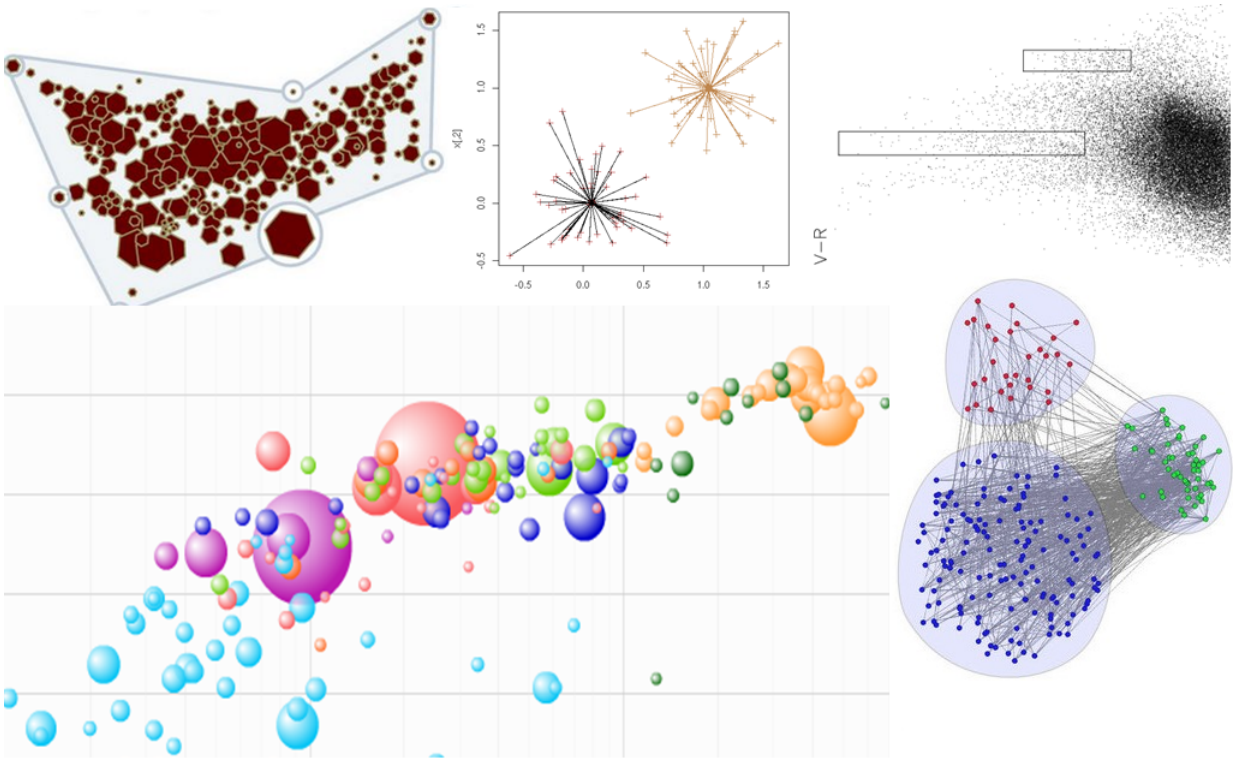


Figura 2.4: Estas gráficas ejemplarizan el método de clustering basado en distancia y en densidad.

Métodos basados en distancia

Los métodos basados en distancia, los outliers se definen como un objeto que esta por lo menos a una distancia d_{min} de porción k de objetos en el conjunto de datos. El problema es entonces encontrar la d_{min} apropiada y k tal que los valores extremos se detecten correctamente con un pequeño número de falsos positivos. En este proceso generalmente es necesario el conocimiento del dominio. Esta metodología no es eficiente con grandes conjuntos de datos.

Se tiene en cuenta la definición propuesta por Knorr y Ng [13]:

Definicion 8 *Un punto x en un conjunto de datos es un outlier con respecto a los parámetros k y d, si no más de k puntos en el conjunto de datos están a una distancia d igual o inferior a x.*

Para entender la definición 6, tomamos por ejemplo el parámetro $k = 3$ y la distancia d como se muestra en la figura 2.5, los puntos x_i y x_j, se definen como los valores extremos, por dentro del círculo. Para cada punto se encuentran no más de 3 puntos de los otros. Y x' es un outlier porque ha excedido el número de puntos dentro del círculo para determinados parámetros k y d.

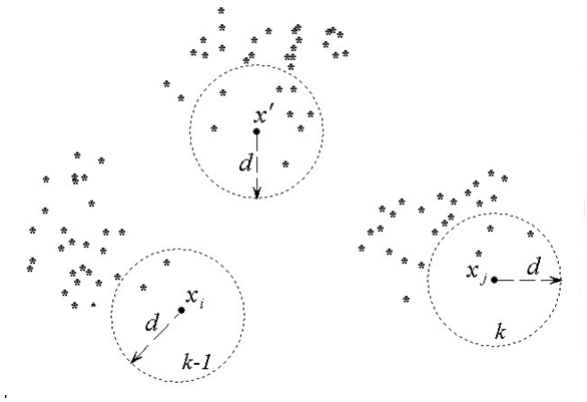


Figura 2.5: Ejemplo de detección de outliers, con métodos de distancia.

Métodos basados en densidad

Estos métodos se enfocan en aplicar un criterio de clúster local. Las agrupaciones son consideradas como regiones densas en el espacio de los datos que están separados por regiones de baja densidad (ruido). Estas regiones pueden tener una forma arbitraria y los puntos dentro de una región pueden ser arbitrariamente distribuidos. Este es robusto ante la presencia de ruido, además es escalable es decir hace un único recorrido del conjunto de datos. Para esto hay una cantidad de algoritmos que realizan este trabajo:

- DBSCAN: Density Based Spatial Clustering of Applications with Noise (Ester et al., KDD'1996).
- OPTICS: Ordering Points To Identify the Clustering Structure (Ankerstet al. SIGMOD'1999).
- DENCLUE: DENsity-basedCLUstEring (Hinneburg y Keim, KDD'1998).
- CLIQUE: Clusteringin QUEst (Agrawalet al., SIGMOD'1998).
- SNN: (SharedNearestNeighbor) density-basedclustering(Ertöz, Steinbach y Kumar, SDM'2003).
- LOF: Local OutlierFactor (Breuning, Kriegel, Ng, Sandre.Sigmond2000). [4], [26]

Ejemplo 9 *Se ha realizado un gráfico en dos dimensiones el cual muestra los datos y el cluster en términos de densidad en diferentes colores. En la figura 2.6, se pueden observar anomalías.*

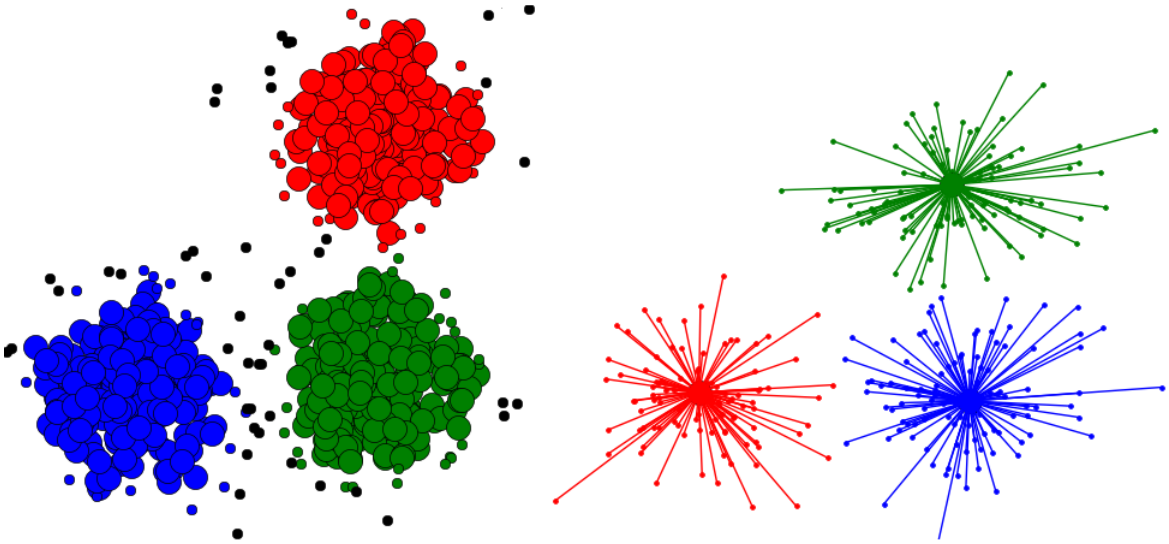


Figura 2.6: Ejemplo de métodos basados en distancia.

Capítulo 3

Metodología propuesta

Sean $X_1, ..., X_n$ realizaciones de la variable objetivo F, de la cual se quiere detectar la presencia de outliers. Se toma $Y_i := X_i - Me$, donde Me es la mediana de los datos.

Ahora se toma $|Y|_{(1)}, ..., |Y|_{(n)}$, la sucesión de los valores absolutos ordenados, se define $|Y_{D_j}| = |Y|_{(j)}$ con $j = 1, ..., n$,

donde D_j es el antirrango de $|Y|_{(j)}$, i.e D_j es el subíndice que tenia originalmente $|Y|_{(j)}$ en la sucesión de valores absolutos $|Y_1|, ..., |Y_n|$.

Sea $\eta_1, ..., \eta_n$ la sucesión dicotomizada, donde en ella se representan las observaciones positivas por unos y las negativas por ceros, de la siguiente manera:

$$\eta_j = \begin{cases} 1, \text{si } Y_{D_j} > 0 \\ 0, \text{en otro caso} \end{cases}$$

La funcion δ_j, donde

$$\delta_j = \begin{cases} 1, \text{si } \eta_j = 1 \\ -1, \text{si } \eta_j = 0 \end{cases}$$

para $j = 1, 2, ..., n$.

Para tomar la cuenta del número de cambios existentes en la sucesión dicotomizada de arriba se definen los siguientes indicadores:

$$I_1 = 1,$$

$$I_j = \begin{cases} 1, \text{si } \eta_{j-1} \neq \eta_j \\ 0, \text{si } \eta_{j-1} = \eta_j \end{cases}$$

donde $j = 2, ..., n$.

Ahora el número de rachas (Una racha es una sucesión ordenada de dos o más simbolos continua de eventos similares o eventos probables) hasta la j-ésima observación en la sucesión dicotomizada se calcula mediante el operador:

$$r_j = \sum_{k=1}^{j} I_k$$

para $j = 1, ..., n$

Obtenemos $r_1, ..., r_n$.

Se clasificarán las observaciones de acuerdo a las siguientes condiciones para las funciones $\xi_i(y_i)$ y $\xi_i^*(y_i)$,

$$\xi_i = \xi_i(y_i) = \begin{cases} 1, \text{si } |y_i| \leq 2\sigma \\ 2, \text{si } 2\sigma < |y_i| \leq 3\sigma \\ 3, \text{si } |y_i| > 3\sigma \end{cases}$$

$$\xi_i^* = \xi_i^*(y_i) = \begin{cases} 1, \text{si } q_1 \leq y_i \leq q_3 \\ 2, \text{si } q_1 - 1,5IQR \leq y_i < q_i \text{ ó } q_3 < y_i \leq q_3 + 1,5IQR \\ 3, \text{si } y_i < q_1 - 1,5IQR \text{ ó } y_i > q_3 + 1,5IQR \end{cases}$$

Estas funciones darán un determinado peso a las observaciones mas "grandes". Para $\xi_i(y_i)$ con el supuesto que la desviación de los datos σ corresponde a la desviación de los datos base, en los cuales no hay presencia de outliers.

NOTA: *los cuartiles q_1 y q_3 deben ser calculados sobre los Y_i ´s o $\xi_i(x_i)$* .

Se tiene un estadístico que cuenta el número de rachas en la sucesión dicotomizada,

$$R = \sum_{k=1}^{n} I_k$$

Finalmente se calcula el estadístico, el cual se define como

$$J = \sum_{i=1}^{n} \frac{\delta_i r_i \xi_i}{r_n},$$

$$J^* = \sum_{i=1}^{n} \frac{\delta_i r_i \xi_i^*}{r_n}.$$

La propuesta se basa en una señal de alerta, el outlier aparecerá como una última observación inmediata, siempre se tiene un J (J^*) base, el cual se calcula con los datos que se tienen en el momento, al realizar un nuevo cálculo con un dato nuevo se obtiene un nuevo J, al cual llamaremos

$$J^R = \sum_{i=1}^{n+1} \frac{\delta_i r_i \xi_i}{r_{n+1}}$$

Notese que no hay un nuevo ordenamiento, no se realiza el tercer paso del proceso cuando entra el nuevo dato, si esto se hiciera las rachas cambiarían, entonces lo que se hace es dejar al final esta observación y determinar si se trata o no de un outlier (según lo estricto del analista), mediante la comparación de $J(J^*)$ y J^R.

Teorema 10 *El estadístico J (J^*) toma valores entre $-2n + \left\lceil \frac{n}{2} \right\rceil - 7$ y $2n - \left\lceil \frac{n}{2} \right\rceil + 7$, y J^R toma valores entre $-2n + \left\lceil \frac{n}{2} \right\rceil - 10$ y $2n - \left\lceil \frac{n}{2} \right\rceil + 10$.*

Demostración 11 *De la definición de R se sabe que $r_j \leq r_n$ esto $\forall j = 1, ..., n$, luego $\frac{1}{r_j} \geq \frac{1}{r_n}$, entonces*

$$J = \sum_{i=1}^{n} \frac{\delta_i r_i \xi_i}{r_n} \leq \sum_{i=1}^{n} \frac{\delta_i r_n \xi_i}{r_n} = \sum_{i=1}^{n} \delta_i \xi_i = \sum_{i=1}^{\lceil \frac{n}{2} \rceil + 1} \delta_i \xi_i + \sum_{i=\lceil \frac{n}{2} \rceil + 2}^{n} \delta_i \xi_i$$

la particion de la sumatoria está basada en la clasificación de los datos para $\xi(\xi^*)$ en virtud del Teorema de Chebyshev (Desigualdad)

$$J \leq \sum_{i=1}^{\lceil \frac{n}{2} \rceil + 1} \delta_i + \sum_{i=\lceil \frac{n}{2} \rceil + 2}^{n} \xi_i \leq \left(\left\lceil \frac{n}{2} \right\rceil + 1 \right) + \left[2 \left(n - \left\lceil \frac{n}{2} \right\rceil \ + 2 + 1 \right) \right] = 2n - \left\lceil \frac{n}{2} \right\rceil + 7$$

y similarmente

$$J \geq \sum_{i=1}^{n} \delta_i \xi_i \geq - \left(\left\lceil \frac{n}{2} \right\rceil + 1 \right) - \left[2 \left(n - \left\lceil \frac{n}{2} \right\rceil \ + 3 \right) \right] = -2n + \left\lceil \frac{n}{2} \right\rceil - 7.$$

Finalmente

$$-2n + \left\lceil \frac{n}{2} \right\rceil - 10 \leq J^R \leq 2n - \left\lceil \frac{n}{2} \right\rceil + 10,$$

Pues $\delta_i = 1$ o -1, y $\xi = 1$ o 2 para a lo más la mitad de los datos y a lo más 3 para el último dato.

□

Criterio de detección

Se puede saber cómo se comportaría el estadístico J^R respecto al estadístico J (J^*).

Sea $J = \sum_{i=1}^{n} \frac{\delta_i r_i \xi_i}{r_n}$, el estadístico base:

El estadístico J^R tomara diferentes valores según el cambio de racha en η_i, donde k toma valores de 1, 2 y 3:

- Cuando η_i pasa de 0 a 00, entonces $J^R = J - k$.
- Cuando η_i pasa de 0 a 01, entonces $J^R = \sum_{i=1}^{n} \frac{\delta_i r_i \xi_i}{r_n+1} + k$.
- Cuando η_i pasa de 1 a 10, entonces $J^R = \sum_{i=1}^{n} \frac{\delta_i r_i \xi_i}{r_n+1} - k$.
- Cuando η_i pasa de 1 a 11, entonces $J^R = J + k$.

Cuando $k = 1$, se dirá que la entrada del dato es una entrada típica en los datos, cuando $k = 2$ se dirá que la entrada del dato es una entrada semitípica en los datos, pero cuando $k = 3$ el dato será declarado como outlier, luego para el caso en que k tome este valor el estadístico dará un valor que mostrará la alerta de entrada de outlier.

Sin importar el valor de la nueva observación a los datos base, al trabajar con las rachas, el cálculo de J^R, siempre ira a tener el mismo valor según el valor de las rachas.

3.1. GLD (Generalized Lambda Distribution) Distribución Lambda Generalizada

En esta sección se indicarán algunas definiciones y generalidades de la GLD [12] que permitirá comparar la potencia de las pruebas propuestas.

Definicion 12 *Distribución Lambda Generalizada.*

La familia de distribuciones λ-generalizada con parámetros λ_1, λ_2, λ_3, λ_4, GLD(λ_1, λ_2, λ_3, λ_4) se define en términos de su función cuantil

$$Q(y) = Q(y, \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{y^{\lambda_3} - (1-y)^{\lambda_4}}{\lambda_2},$$

donde λ_1 y λ_2 son respectivamente parámetros de localización y escala, mientras λ_3 y λ_4 determinan la falta de simetría y curtosis de GLD.

Teorema 13 *GLD(λ_1, λ_2, λ_3, λ_4) tiene como función de densidad a*

$$f(x) = \frac{\lambda_2}{\lambda_3 y^{\lambda_3 - 1} + \lambda_4 (1-y)^{\lambda_4 - 1}},$$

donde $x = Q(y)$.

No todas las $Q(x)$ proporcionan distribuciones válidas para GLD(λ_1, λ_2, λ_3, λ_4). Es necesario que $f(x) \geq 0$ y $\int_{-\infty}^{\infty} f(x)dx = 1$. Tambien hay condiciones para los parametros, en este caso para los signos y valores específicos que pueden tomar, por ejemplo hay condiciones para λ_3 y λ_4 que están en regiones del plano cartesiano dadas por unas curvas definidas.

Ejemplo 14 *Coeficiente de fricción de un metal*

La gráfica de $f(x)$, función de densidad, se sigue considerando GLD(λ_1, λ_2, λ_3, λ_4) con parámetros λ_1 =0.0305, λ_2 =1.3673, λ_3 =0.004581, λ_4 =0.01020, se tiene la función GLD(λ_1, λ_2, λ_3, λ_4) que se sigue de la función cuantil así [12]:

$$Q(y) = 0{,}0305 + \frac{y^{0{,}004581} - (1 - y^{0{,}01020})}{1{,}3673}.$$

Se encuentra que para y =0.25 la función cuantil Q(0.25)=0.028013029, x =0.028013029, ahora usando el Teorema 13 con los valores $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ se obtiene

$$f(0{,}028013029) = 43{,}0399612$$

Por tanto (0.028013029,43.0399612) será uno de los puntos de la grafica $f(x)$, procediendo de esta manera para 0.01,...,0.99 (1 %,...,99 %) se obtiene la figura 3.1.

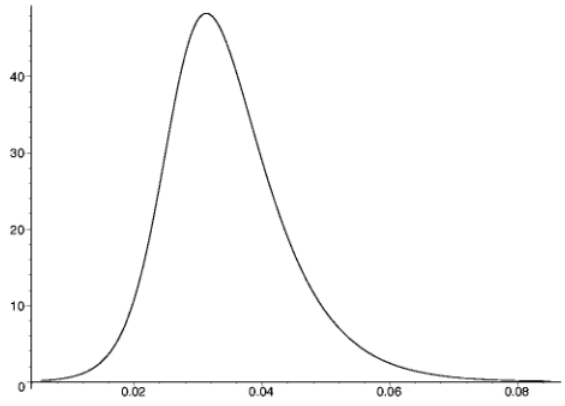


Figura 3.1: fdp para GLD(0.0305, 1.3673, 0.004581, 0.01020).

3.2. Montecarlo

Llamada así en honor del casino Montecarlo (Principado de Mónaco) por ser capital de los juegos de azar y por ser la ruleta generadora de números aleatorios. El desarrollo de la metodología Monte Carlo data desde 1944, de trabajos realizados en el desarrollo de la bomba atómica en

la II guerra mundial en el laboratorio nacional de Álamos EEUU por John Von Newmann y Sranislaw Ulam. Este trabajo conlleva una simulación de problemas de probabilidad.

Métodos Montecarlo: Técnica que puede ser usada para resolver un problema matemático o estadístico.

Simulación Montecarlo: [10] Representación fictisia de la realidad, usa muestras repetidas para determinar las propiedades de algún fenómeno.
La simulación Montecarlo es una técnica cuantitativa que hace uso de la probabilidad y los ordenadores para simular situaciones, mediante modelos matemáticos. La clave para la simulación MC consiste en crear un modelo, el cual dirá el comportamiento global del sistema pensando siempre en las variables a estudiar, una vez identificadas las variables, el experimento se lleva a cabo de la siguiente manera:

1. Se generan muestras aleatorias con la ayuda del ordenador.

2. Se analiza el comportamiento del sistema ante los valores generado.

Tras repetir el proceso n veces dispondremos de n observaciones sobre el comportamiento del modelo, lo cual nos será de utilidad para entender el funcionamiento del mismo, así el análisis será de mayor precisión a medida que el número de repeticiones n aumenta.

3.3. Comportamiento de la metodología propuesta

Se presenta la comparación del comportamiento de las pruebas por simulación de Monte Carlo y se generan 5 distribuciones pertenecientes a la GLD, esta simulación lleva a indicar que la propuesta es más eficiente. Se evalúan los siguientes estadísticos:

$$J = \sum_{i=1}^{n} \frac{\delta_i r_i \xi_i}{r_n},$$

$$J^* = \sum_{i=1}^{n} \frac{\delta_i r_i \xi_i^*}{r_n}.$$

Las 5 distribuciones continuas generadas por la DLG que se muestran en el cuadro 3.1.

Distribución	$\lambda_1, \lambda_2, \lambda_3, \lambda_4$
$N(0,1)$	(0,0.1975,01349,0.1349)
Exponencial ($\theta = 1$)	(0.006862,-0.0010805,-0,4072$x10^{-5}$,-0.001076)
Uniforme $(0,1)$	(0.5,2,1,1)
Weibull ($\alpha = 1; \beta = 5$)	(0.9935, 1.0491, 0.2121, 0.1061)
Chi-S ($v = 3$)	(0.8596, 0.0095443, 0.002058, 0.02300)

Cuadro 3.1: Parametros de DLG.

Para estimar la potencia de los estadísticos se realizo un programa en SAS. El algoritmo que se siguio es el siguiente.

1. Se selecciona una muestra aleatoria $u_1, ..., u_n$ de la distribución $U(0, 1)$.

2. Se transforman la muestra $u_1, ...u_n$ en una sucesión $x_1^*, ..., x_n^*$ utilizando la función cuantil de DLG, así:

$$Q(y) = x_i^* = \lambda_1 + \frac{u_i^{\lambda_3} - (1 - u_i)^{\lambda_4}}{\lambda_2}$$

con $i = 1, ..., n$.

La sucesión $x_1^*, ..., x_n^*$ es muestra aleatoria de una DLG con parámetros $\lambda_1, \lambda_2, \lambda_3, \lambda_4$.

3. Se transforma $x_i = x_i^* - \theta$ para que la distribución $x_1^*, ..., x_n^*$ tenga mediana cero, donde

$$\theta = \lambda_1 + \frac{0,5^{\lambda_3} - 0,5^{\lambda_4}}{\lambda_2}$$

4. Se calculan los valores de los estadísticos que se van a comparar usando las observaciones de la muestra $x_1, ..., x_n$.

5. Se realizan las respectivas pruebas y para cada uno se determina si se declara el dato entrante como outlier, aleatorizando la prueba.

6. Se aplica el anterior proceso 1000 veces, y se estimar el comportamiento de cada una de las pruebas, así:

$$\pi^* = \frac{\text{Número de detecciones}}{\text{Número de outliers}}$$

Capítulo 4

Aplicaciones

Empresas de telefonía, bancos y bolsa utilizan métodos estáticos para la detección de outliers. A continuación se muestra un ejemplo con las diferentes metodologías usadas y con la metodología propuesta.

4.1. Bolsa

La detección correcta y oportuna de estos outliers en la bolsa se traduce en ganancias o pérdidas millonarias. La figura 4.1 muestra los precios de cierre de una petrolera entre el 3 de enero y el 15 de abril del 2011.

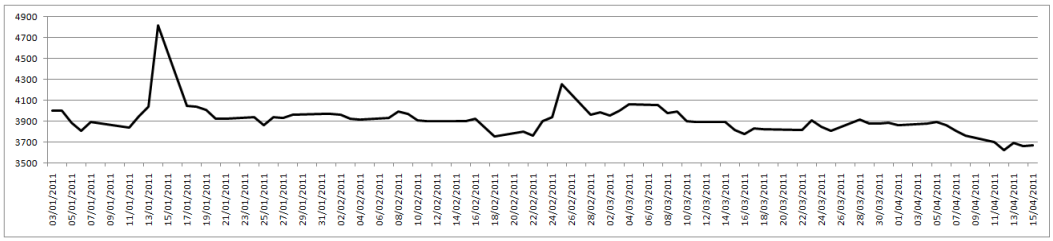


Figura 4.1: Precio de cierre de Ecopetrol.

Los dos primeros métodos que se muestran son métodos basados en estadísticas, que se fundamentan en intervalos de confianza. El primer método toma todo el histórico que tiene del precio de cierre y calcula un intervalo de confianza en términos de desviaciones estándar de la siguiente manera:

Se supone X como la v.a que describe el precio de cierre, el intervalo es

$$[\bar{x} - 3\sigma, \bar{x} + 3\sigma]$$

Los cálculos para estos estimadores son $\bar{x}$ =3910.753425 y σ =148.9049642 (donde $\bar{x}$ es el promedio de todos los datos), así se obtiene:

[3464.038532,4357.468317]. Vease figura 4.2.

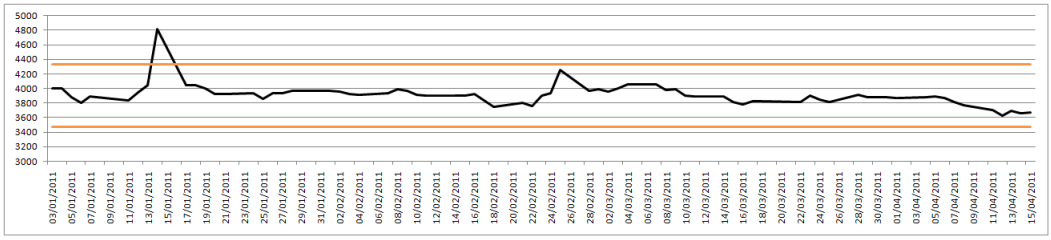


Figura 4.2: Intervalo de confianza.

El intervalo detectó uno de los tres outliers.

El otro método está basado en el mismo intervalo, pero el promedio se halla en los días hábiles de la semana, la figura 4.3 representa este proceso.

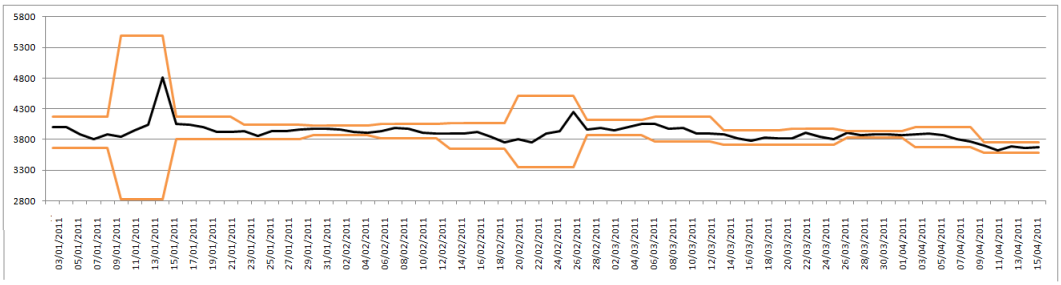


Figura 4.3: Intervalo de confianza 2.

En este caso ninguno de los intervalos detecto los outliers. Un problema de estos métodos es que los outliers existentes en los datos modifican los intervalos de confianza. Otro problema es que supone que los datos se distribuyen normalmente con media 0 y varianza 1 sin ser así. Estos dos métodos no son más que la aplicación de la región de outlier.

El tercer método es el boxplot, se representa por la figura 4.4

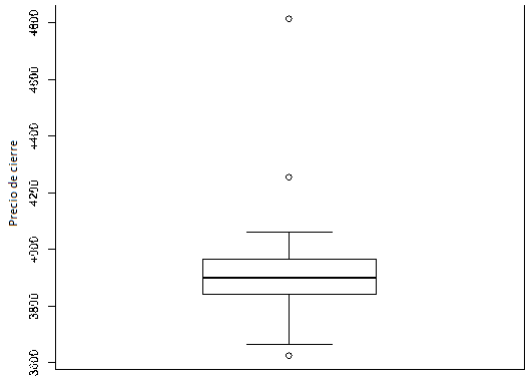


Figura 4.4: Boxplot

El boxplot detecto los tres outliers. El problema es que estamos tratando con datos dinamicos, y el boxplot funciona bien para datos estaticos.

El programa de SAS del método de detección temprana propuesto se corrió comenzando con $n = 5$ (con J). Cuando el programa se encontraba en $n = 9$ se encontró un outlier, esto se concluyo ya que $J = 0,8$, mientras que $J^R = 3,8$ ($4815), luego se reemplazo este valor por la mediana y se siguió haciendo el proceso. Se encontró otro outlier cuando el proceso se encontraba en $n = 39$ donde $J = -1,9$ y $J^R = 1,19047619$ ($4255), de nuevo se reemplazo este valor por la mediana y se siguió haciendo el proceso. En $n = 70$ se encontró otro outlier, esto se concluyo porque $J = 17$ y $J^R = 14$ ($3625). El método propuesto detectó estos outliers como se muestra en la figura 4.5.

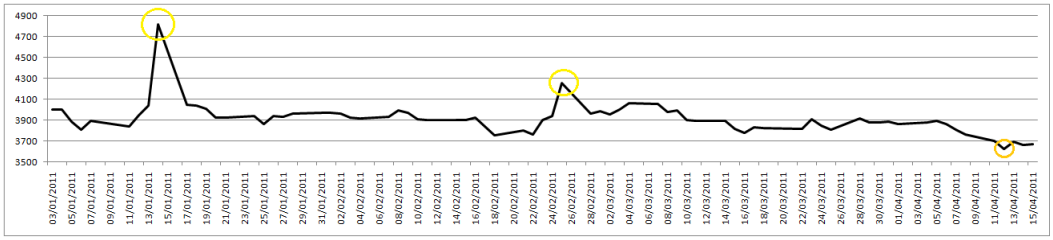


Figura 4.5: Método de detección temprana.

Capítulo 5

Resultados y conclusiones

Los estadísticos J y J^* se evaluaron para $n = 10$ y $n = 100$, con 1000 réplicas, los outliers que fueron insertados en los datos de las diferentes distribuciones fueron tomados en el mínimo umbral de cada una.

Para el caso $n = 10$ se puso un outlier en diferentes posiciones con la condición de que fuera mayor que 5. En el cuadro 5.1 se puede ver como el estadístico J funcionó mejor que J^*, debido a que en la distribución exponencial J^* encontró un falso positivo y en la distribución uniforme no detecto el outlier.

Distribución	$n = 10$	
	J	J^*
$N(0, 1)$	1/1	1/1
Exponencial	1/1	2/1
Uniforme	1/1	0/1
Weibull	1/1	1/1
Chi-S	1/1	1/1
Aleatorio(1000-2000)	1/1	1/1

Cuadro 5.1: Estimación del comportamieno de las pruebas $n = 10$.

Para el caso $n = 100$ se pusieron 5 outliers en diferentes posiciones, en este caso en rangos de 20, es decir el primer outlier estaba en los primeros 20 datos, el segundo entre los 20 y 40 datos, el tercero entre los 40 y 60 datos, y así sucesivamente. De nuevo con la misma condición de que estos estuvieran en una posición mayor que 5. El cuadro 5.2 muestra que J^* tiene fallos en las diferentes distribuciones, teniendo falsos positivos y outliers no detectados. El J tuvo un mejor comportamiento, pero si hubo casos de falsos positivos como fue el caso de la distribución exponencial y la Weibull.

Distribución	$n = 100$									
Outliers	1^o	2^o	3^o	4^o	5^o	1^o	2^o	3^o	4^o	5^o
	J					J^*				
N(0,1)	1/1	1/1	1/1	1/1	1/1	2/1	2/1	2/1	2/1	2/1
Exponencial	1/1	1/1	1/1	3/1	2/1	1/1	1/1	3/1	3/1	3/1
Uniforme	1/1	1/1	1/1	1/1	1/1	0/1	0/1	0/1	0/1	1/1
Weibull	1/1	1/1	1/1	2/1	2/1	1/1	1/1	4/1	4/1	3/1
Chi-S	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
Aleatorio(1000-2000)	1/1	1/1	1/1	1/1	1/1	0/1	1/1	1/1	1/1	1/1

Cuadro 5.2: Estimación del comportamieno de las pruebas $n = 100$

A medida que n aumenta, debido a lo aleatorio de los datos ambos estadísticos dectarón falsos positivos. Esto lleva a proponer aplicar el método con una muestra recortada de entre 50 y 100 datos para renovar la entrada de los posibles datos.

Para los tamaños de muestra tomados se concluye que el estadístico J fue mejor que J^*, no solo por detección, si no también por los valores que este toma. Los valores de J son muchas veces menores que los de J^*. A medida que se aumente los datos base, es posible que ambas pruebas se vean afectadas por la detección de falsos positivos, es recomendable tomar muestras recortadas.

Bibliografía

[1] Acuna E., Rodriguez C. A, *Meta analysis study of outlier detection methods in classification* . Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, Retrived from academic.uprm.edu/ eacuna/ paperout.pdf. In proceedings IPSI 2004, Venice.

[2] Caussinus H., Roiz A, *Interesting projections of multidimensional data by means of generalized component analysis.* In Compstat 90, 121-126, Heidelberg: Physica, 1990.

[3] Davies L., Gather U, *The identication of multiple outliers.* Journal of the American Statistical Association, 88(423), 782-792, 1993.

[4] Fernando Berzal, *Notas de clase Clustering.* Departamento de ciencias de la computación e I.A. Universidad de Granada.

[5] Giovany Babativa, Jimmy Corzo, *Propuesta de una prueba de rachas recortadas para hipótesis de simetría* Revista Colombiana de Estadística. Volumen 33, No 2, pp.251 a 271. Diciembre de 2010.

[6] Hampel F. R., *A general qualitative definition of robustness.* Annals of Mathematics Statistics. 42, 1887-1896, 1971.

[7] Hampel F. R., *The influence curve and its role in robust estimation.* Journal of the American Statistical Association, 69, 382-393, 1974.

[8] Irad Ben-Gal, *Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers.* Academic Publishers, ISBN 0-387-24435-2. 2005.

[9] Ismael Sánchez , *notas de clase, Métodos estadísticos para la mejora de la calidad* Departamento de Estadística y Econometría Universidad Carlos III de Madrid. 2002.

[10] Javier Faulin, Angel Juan, *Notas de clase Simulación de Montecarlo.*

[11] Josep Maria Puigvert Gutierrez and Josep Fortiana Gregori, *Cluster techniques applied to outlier detection of financial market series using a moving window filtering algorithm.* European Central Bank. 2008.

[12] Zaven A. Karian, Edward J. Dudewicz, *Handbook of Fitting Statistical Distributions with R-Zaven A* CRC Press. 2011

[13] Knorr E., Ng. R., *Algorithms for mining distance-based outliers in large datasets.* In Proc. 24th Int. Conf. Very Large Data Bases (VLDB), 392-403, 24-27, 1998.

[14] Liu H., Shah S., Jiang W., *On-line outlier detection and data cleaning.* Computers and Chemical Engineering, 28, 1635-1647, 2004.

[15] Mahalanobis, Prasanta Chandra, *On the generalised distance in statistics.* Proceedings of the National Institute of Sciences of India 2. 49-55. Retrieved 2012-05-03. (1936)

[16] Martin R. D., Thomson D. J., *Robust-resistant spectrum estimation.* In Proceeding of the IEEE, 70, 1097-1115, 1982.

[17] Milton Januario Rueda Varon Ph.D., *A nonparametric test based on runs for a single sample location problem* Universidad Konstanz Alemania. 2010.

[18] Miodrag Lovric, *International encyclopedia of statical science.* Springer 2010.

[19] Osborne, JW., Overbay, A., *The power of outliers (and why researchers should always check for them).* Practical Assessment, Research & Evaluation, 9(6). 2004.

[20] Rousseeuw P., *Multivariate estimation with high breakdown point.* In: W. Grossmann et al., editors, Mathematical Statistics and Applications, Vol. B, 283-297, Akademiai Kiado: Budapest, 1985.

[21] Rousseeuw P., Leory A., *Robust Regression and Outlier Detection* Wiley Series in Probability and Statistics, 1987.

[22] Shewart Page, E. S. *Continuous Inspection Scheme* Math. Annalen **70** (1911), 351–376. Biometrika 41 (1/2): 100-115. (June, 1954) JSTOR 2333009.

[23] Songwon Seo, *A review and comparison of methods for detecting outliers in univariate data.* University of pittsburgh. 2006.

[24] Tukey, John W., *Exploratory Data Analysis.* Addison-Wesley. (1977) ISBN 0-201-07616-0. OCLC 3058187.

[25] Peter J. R ousseeuw and Katrien Van Driessen, *A fast Algorithm for the Minimum Covariance Determinant Estimator.* Technometrics , August 1999, vol. 41, No 3

[26] http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/Clustering/