Direcciones de Proyección Aleatorias Específicas para Desviaciones Ajustadas por Asimetría

José Schneider Londoño González

Trabajo de grado para optar al título de
Magíster en Ciencia de Datos

Director:
Santiago Ortiz Arias

FACULTAD DE INGENIERÍA, DISEÑO Y CIENCIAS APLICADAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2024

Resumen

Los estimadores basados en métodos estadísticos suelen ser afectados por datos asimétricos, razón por la cual la detección de atípicos con estas técnicas en presencia de distribuciones asimétricas no suele ser eficaces. Este trabajo presenta una nueva propuesta del procedimiento Skewness Adjusted Outlyingness (SAO), que es un método de búsqueda de proyecciones para la detección de valores atípicos multivariados. Estas direcciones se obtienen a través de la dirección que maximiza el tercer momento de muestra al cuadrado de los datos proyectados, que luego se utiliza como semilla para calcular $4p^2$ direcciones específicas obtenidas de un muestreo estratificado. Se compara el rendimiento de la versión SAO propuesta con el SAO estándar. Los resultados empíricos, tanto en términos de la tasa de verdaderos positivos como de la tasa de falsos positivos, muestran que el método propuesto es resistente a la hora de detectar valores atípicos multivariados.

Índice de Contenidos

Índice de Tablas

Índice de Figuras

1. Introducción

En el análisis de datos, la detección de valores atípicos es crucial debido a su capacidad para desviar los resultados estadísticos y llevar a conclusiones erróneas. Los valores atípicos son puntos que no encajan con el resto de los datos y pueden tener un impacto significativo en el análisis. Estos puntos extremos, conocidos como valores atípicos, deben ser identificados y manejados adecuadamente para evitar interpretaciones erróneas de los resultados. Este documento explora el concepto de valores atípicos, su importancia y los métodos para detectarlos, además de presentar una nueva propuesta para mejorar la precisión y eficiencia en su identificación.

Uno de los métodos para la detección de valores atípicos es el algoritmo Stahel-Donoho (SD), propuesto por Stahel (1981) y Donoho (1982). Este método robusto y equivariante afín calcula aleatoriamente un número fijo de direcciones y calcula la medida de atipicidad en cada dirección para los datos proyectados. A pesar de su alto punto de quiebre y eficiencia asintótica, el algoritmo SD presenta deficiencias significativas, como el alto esfuerzo computacional requerido y la posibilidad de que las direcciones aleatorias no sean útiles, especialmente cuando la dimensión del espacio muestral aumenta (Maronna y Yohai, 1995; Gervini, 2002). Estas limitaciones motivan la búsqueda de métodos alternativos que mantengan la robustez del SD pero mejoren su eficiencia y utilidad en espacios de alta dimensión.

Otro enfoque es el método Skewness Adjusted Outlyingness (SAO), propuesto por Hubert y Van der Veeken (2008). Este método se basa en la asimetría multivariada para identificar valores atípicos, ajustando la medida de atipicidad para adaptarse a la asimetría inherente de los datos. La ventaja de SAO es su sensibilidad a los valores atípicos en distribuciones sesgadas, lo que lo convierte en una herramienta versátil para diversos tipos de datos (Loperfido, 2018). Sin embargo, SAO también tiene sus limitaciones, como la necesidad de generar un gran número de direcciones aleatorias $250p$, lo que puede ser ineficaz y no garantiza una detección exitosa en todos los casos.

En este trabajo, proponemos una alternativa al método SAO que aborda sus principales limitaciones. En lugar de generar direcciones aleatorias, definimos un conjunto de direcciones de proyección más informativas calculadas mediante un procedimiento de muestreo estratificado sobre una dirección semilla. Esta metodología reduce significativamente el esfuerzo computacional y mejora la precisión en la detección de valores atípicos multivariados. El documento describe el estimador propuesto y su derivación, presenta un estudio de simulaciones con diferentes escenarios de contaminación y ofrece algunas conclusiones. Esta nueva aproximación promete mejorar la robustez y eficiencia en la identificación de valores atípicos, contribuyendo así a la integridad del análisis de datos.

2. Contexto y Antecedentes

Los estimadores basados en métodos estadísticos tradicionales suelen verse afectados por datos asimétricos, lo que dificulta la detección eficaz de valores atípicos en distribuciones asimétricas. Esta limitación subraya la necesidad de contrastar y evaluar diversas herramientas de detección de atípicos para identificar técnicas más efectivas y robustas. Encontrar métodos que generen hallazgos relevantes es crucial para el correcto uso y aplicación de técnicas estadísticas en entornos investigativos y productivos. Por esta razón, la investigación en la mejora de técnicas de detección de atípicos es de gran valor, ya que puede contribuir significativamente a la precisión y fiabilidad del análisis de datos.

Una de las técnicas más conocidas para la detección de valores atípicos en datos multivariados es la búsqueda de proyecciones. Este enfoque implica proyectar los datos en varias direcciones y analizar las propiedades estadísticas de estas proyecciones para identificar puntos que se desvían notablemente del resto. Dentro de este marco, el algoritmo Stahel-Donoho (SD) se destaca como un método robusto y eficiente para la detección de atípicos multivariados. El algoritmo SD calcula aleatoriamente un número fijo de direcciones y evalúa la medida de atipicidad en cada una de ellas para los datos proyectados.

El algoritmo SD es equivariante afín (es decir, el resultado de la detección de valores atípicos no cambia si los datos se rotan, escalan o trasladan) y tiene un alto punto de quiebre (es decir, tolera un gran porcentaje de valores atípicos sin que sus resultados se vean significativamente afectados), lo que lo hace resistente a la influencia de valores atípicos extremos. Además, su eficiencia asintótica ha sido confirmada por estudios de su función de influencia (Gervini, 2002). Sin embargo, a pesar de sus ventajas, el algoritmo SD presenta ciertas limitaciones significativas. Entre ellas se incluyen el alto esfuerzo computacional necesario debido a la gran cantidad de direcciones que deben evaluarse y la posibilidad de que las direcciones aleatorias no siempre sean útiles, especialmente a medida que aumenta la dimensión del espacio muestral. Estas limitaciones motivan la búsqueda de mejoras y alternativas al método original.

La evolución y refinamiento de las técnicas de detección de atípicos han llevado al desarrollo de metodologías como el método Skewness Adjusted Outlyingness (SAO), propuesto por Hubert y Van der Veeken (2008). El método SAO se basa en la asimetría multivariada y ajusta la medida de atipicidad para adaptarse mejor a la asimetría inherente de los datos, haciéndolo más sensible a valores atípicos en distribuciones sesgadas (Loperfido, 2018). Aunque el método SAO presenta ventajas en términos de sensibilidad y versatilidad, también enfrenta desafíos, como la necesidad de generar un gran número de direcciones aleatorias, lo que puede resultar ineficiente y no garantizar una detección exitosa en todos los casos. Estos antecedentes destacan la importancia de continuar investigando y desarrollando nuevas metodologías que superen las limitaciones existentes, mejorando la detección de valores atípicos en diversas aplicaciones.

3. Planteamiento del Problema y Justificación

En el análisis de datos multivariados, la detección de valores atípicos es un desafío crítico que puede influir significativamente en la interpretación y resultados de los estudios estadísticos. Los valores atípicos pueden desviar los análisis, llevar a conclusiones erróneas y afectar la calidad de las decisiones basadas en dichos análisis. Los métodos tradicionales, como el algoritmo Stahel-Donoho (SD) y el método Skewness Adjusted Outlyingness (SAO), han sido desarrollados para abordar este problema. Sin embargo, ambos métodos presentan limitaciones que afectan su eficacia y aplicabilidad. El algoritmo SD, aunque robusto y eficiente, requiere un alto esfuerzo computacional debido a la necesidad de calcular numerosas direcciones aleatorias, las cuales no siempre son útiles. Por otro lado, el método SAO, aunque es más sensible a valores atípicos en distribuciones sesgadas, también depende de un gran número de direcciones aleatorias y no garantiza una detección exitosa en todos los casos. Estas limitaciones subrayan la necesidad de desarrollar métodos más eficientes y precisos para la detección de valores atípicos multivariados.

La justificación de esta investigación radica en la creciente necesidad de métodos de detección de valores atípicos que sean no solo robustos, sino también computacionalmente eficientes y adaptables a diferentes estructuras de datos. La detección precisa de valores atípicos es esencial en numerosos campos, incluyendo la ciencia de datos, la economía, la biología y la ingeniería, donde las decisiones críticas dependen de la calidad y precisión del análisis de datos. La propuesta de una alternativa al método SAO, basada en direcciones de proyección más informativas calculadas mediante un procedimiento de muestreo estratificado, promete abordar las limitaciones actuales y ofrecer un método más efectivo. Este enfoque no solo reducirá el esfuerzo computacional, sino que también mejorará la precisión de la detección de valores atípicos, contribuyendo significativamente a la fiabilidad de los análisis en contextos multivariados.

La relevancia de esta investigación se extiende más allá de la teoría estadística, impactando directamente en la práctica de análisis de datos en diversas disciplinas. Al mejorar la capacidad para detectar valores atípicos, se potencia la integridad de los datos y la validez de los análisis subsecuentes, lo cual es crucial para obtener resultados confiables y tomar decisiones informadas. La metodología propuesta no solo tiene el potencial de ser una herramienta valiosa para los investigadores, sino también para los profesionales que manejan grandes volúmenes de datos en sus respectivas áreas. La aplicabilidad de un método de detección de valores atípicos más eficiente y preciso puede transformar prácticas en sectores como la salud, la finanza y la manufactura, donde la precisión de los datos es de suma importancia. Por tanto, esta investigación no solo contribuye al avance del conocimiento en técnicas estadísticas, sino que también ofrece soluciones prácticas a problemas reales en el análisis de datos multivariados.

4. Objetivos del Proyecto

4.1. Objetivo General

- Diseñar una metodología de generación de vectores aleatorios-específicos en métodos de proyecciones por atipicidad de tipo asimétrico, para detección de atípicos multivariantes.

4.2. Objetivos Específicos

- Examinar las metodologías de proyecciones actuales de detección de valores atípicos.

- Desarrollar un procedimiento para la generación de vectores de proyección informativos.

- Implementar y evaluar la nueva metodología en diferentes escenarios de datos.

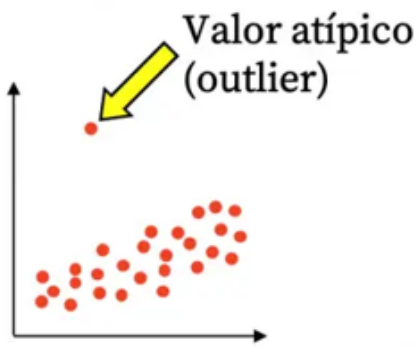
5. Marco Teórico

5.1. Dominio del Problema.

5.1.1. Datos Atípicos.

Los datos atípicos o outliers se refieren a observaciones que se apartan significativamente de la tendencia general o el patrón de comportamiento esperado en un conjunto de datos. Estas observaciones atípicas pueden ser valores inusuales o extremos en comparación con el resto de los datos.

Figura No. 1. Diagrama de Dispersión Bivariante con Outlier.



Fuente: Martín, B. de U. (2022). *Los valores atípicos, anómalos o Outliers en estadística: ¿Qué son, por qué aparecen y cómo controlarlos?*. Tomado de **https://trabajofinal.es/valores-atipicos-outliers-estadistica/#google_vignette**

En cuanto a las causas de la presencia de outliers en los datos, éstas pueden variar y pueden ser el resultado de diferentes factores como, por ejemplo, los errores de medición, en este caso los outliers pueden surgir debido a errores en el proceso de medición o recopilación de datos, ya sea por fallas técnicas o humanas. También pueden ser originados por eventos inusuales: Algunos outliers pueden ser el resultado de eventos inesperados o poco comunes que ocurren durante la recolección de datos. Estos eventos pueden influir en las mediciones y generar valores atípicos. Finalmente, también pueden ser generados por fenómenos extremos en el contexto de estudio. Por ejemplo, en datos climáticos, un valor de temperatura extremadamente alto o bajo puede considerarse un outlier.

Es fundamental considerar que los outliers pueden tener impactos negativos significativos en el análisis estadístico y los resultados de la modelación. La presencia de outliers puede sesgar las estimaciones de parámetros estadísticos, como la media o la desviación estándar, conduciendo a resultados incorrectos o poco representativos. Además, los outliers pueden afectar la validez de las pruebas de hipótesis y las conclusiones estadísticas; incluso un único outlier puede alterar los resultados y llevar a conclusiones incorrectas. Si no se manejan adecuadamente, los outliers pueden distorsionar los modelos predictivos, lo que puede resultar en predicciones inexactas o poco confiables. Asimismo, si los outliers no se identifican y gestionan correctamente, los resultados y conclusiones obtenidos de los datos pueden no ser generalizables ni aplicables al contexto de estudio.

Por lo tanto, la detección y el manejo adecuado de los outliers son aspectos cruciales en la investigación estadística. El uso de técnicas robustas de detección de outliers es esencial para mitigar los efectos negativos que estos pueden tener en el análisis y los resultados del modelo. La identificación y corrección de outliers no solo mejora la precisión de los análisis estadísticos, sino que también asegura que las inferencias y predicciones derivadas de los modelos sean válidas y aplicables a la población de interés. En resumen, una gestión eficaz de los outliers es indispensable para garantizar la fiabilidad y generalización de los hallazgos en cualquier estudio estadístico.

5.1.2. Datos Asimétricos.

La asimetría de los datos se refiere a la falta de simetría en la distribución de los valores en un conjunto de datos. Una distribución simétrica es aquella en la que los valores están igualmente distribuidos a ambos lados de su media, mientras que una distribución asimétrica presenta una tendencia hacia un lado en particular.

La presencia de asimetría en los datos puede tener implicaciones en la detección de outliers, ya que los outliers pueden afectar la forma de la distribución y provocar un sesgo en las medidas de tendencia central, como la media. En una distribución asimétrica, los outliers pueden estar más influenciados por los valores extremos en el lado de la cola larga de la distribución, lo que puede dificultar su identificación.

Figura No. 2. Gráfico Tipos de Asimetría.

Fuente: belenstgo.bigpress.net. (2019). Asimetría. Tomado de https://belenstgo.bigpress.net/texto-diario/mostrar/1609381/asimetria

Existen diversos métodos y técnicas para identificar y tratar la asimetría en los datos, entre los que se incluyen la transformación de datos, la cual es una estrategia común para tratar la asimetría que consiste en aplicar transformaciones a los datos, como la transformación logarítmica, la raíz cuadrada o la transformación Box-Cox. Estas transformaciones pueden ayudar a reducir la asimetría y hacer que los datos se ajusten mejor a los supuestos de normalidad. También están las medidas de asimetría como el coeficiente de asimetría de Pearson o el coeficiente de asimetría de Fisher, proporcionan una cuantificación de la asimetría en una distribución. Estas medidas pueden utilizarse para evaluar el grado de asimetría en los datos y tomar decisiones sobre el tratamiento de la asimetría.

La presencia de asimetría en los datos puede influir en la elección y el rendimiento de las técnicas robustas de detección de outliers. Se puede usar resistencia a la asimetría, las cuales son técnicas robustas que están diseñadas para ser resistentes a los valores atípicos y los efectos de la asimetría. Estas técnicas pueden ser más efectivas en la detección de outliers en presencia de datos asimétricos, ya que se basan en medidas de centralidad y dispersión más robustas a las desviaciones extremas. También se pueden usar técnicas con sensibilidad a la asimetría, estas técnicas robustas pueden ser más o menos sensibles a la asimetría en los datos. Por ejemplo, las técnicas basadas en medianas pueden ser más efectivas en distribuciones asimétricas con colas largas, mientras que las técnicas basadas en media pueden ser más sensibles a los valores extremos en una cola larga.

5.1.3. Estimadores.

En cuanto a los estimadores, en el contexto de detección de outliers con técnicas robustas, los estimadores robustos son aquellos que están diseñados para ser menos influenciados por los valores atípicos y los datos no normales. Estos estimadores, como la mediana y la media recortada, son menos sensibles a los outliers y pueden proporcionar una mejor estimación de los parámetros de interés en presencia de datos atípicos. Los estimadores robustos son utilizados en las técnicas robustas de detección de outliers para calcular medidas de centralidad y dispersión que son menos afectadas por los valores extremos.

5.2. Dominio de la Solución

5.2.1. Estimador Stahel-Donoho (SD)

El Estimador Stahel-Donoho (SD) es un estimador robusto empleado en el análisis de datos para calcular una medida de tendencia central resistente a los outliers y la presencia de datos asimétricos. Fue propuesto por los estadísticos Stahel y Donoho como una alternativa a los estimadores convencionales, que son sensibles a los valores atípicos y a las distribuciones no normales. Este estimador se basa en el principio de minimización de la función de riesgo influenciada por los outliers, lo que lo hace adecuado para la estimación en presencia de datos contaminados o distribuciones asimétricas. En lugar de calcular una sola medida de tendencia central, el SD proporciona un conjunto de estimaciones conocidas como *centros afinadores* (*hubers* en inglés), que se adaptan a diferentes proporciones de datos contaminados y niveles de asimetría.

El estimador SD es un estimador robusto de la localización y dispersión multivariada. Se define como una media ponderada y una matriz de covarianza, donde los pesos se basan en una medida de atipicidad calculada mediante una función de penalización. La medida de atipicidad se basa en el máximo de la proyección unidimensional en la cual la observación es más atípica, para todas las direcciones de proyección posibles. Luego, los pesos se utilizan para reducir el peso de las observaciones más atípicas.

Considerando la muestra multivariada $\mathbf{x} = (x_{1,} \dots, x_n)$, y $S_d = \{\boldsymbol{d} \in \mathbb{R}^p : \boldsymbol{d}'\boldsymbol{d} = 1\}$ el conjunto de todas las direcciones unitarias de proyección p-dimensional. La atipicidad SD $r(\cdot)$ de un punto de datos x_i en una dirección $\boldsymbol{d} \in S_d$ típicamente se calcula como la distancia entre las observaciones proyectadas $\boldsymbol{d}'x_i$ y una estimación de ubicación

univariada $\mu(\cdot)$ re-escalada por una estimación de dispersión univariada $\sigma(\cdot)$. Por lo tanto, para cualquier x_i, el $r(\, x_i, \mathbf{x}) \equiv r_i$, se define como:

$$r(\, x_i, \mathbf{x}) \equiv \sup_{\boldsymbol{d} \in S_d} \frac{|\boldsymbol{d}' x_i \, - \mu(\boldsymbol{d}' \mathbf{x})|}{\sigma(\boldsymbol{d}' \mathbf{x})}$$

Para hacer $r(\cdot)$ una medida robusta, $\mu(\cdot)$ y $\sigma(\cdot)$ suelen ser la mediana muestral y las estadísticas MAD, respectivamente (Stahel, 1981; Donoho, 1982). Los valores grandes de atipicidad indican qué puntos son particularmente atípicos en relación con el resto de los datos principales, mientras que un valor de atipicidad cercano a 0 indica que el punto está cerca de la mediana y, por lo tanto, no es atípico. Así, el estimador robusto SD para la localización y dispersión multivariada se define como:

$$\hat{\boldsymbol{\mu}}_{SD} = \frac{\sum_{1=1}^{n} \omega_i x_i}{\sum_{1=1}^{n} \omega_i} \text{ y } \hat{\mathbf{S}}_{SD} = \frac{\sum_{1=1}^{n} \omega_i (x_i - \hat{\boldsymbol{\mu}}_{SD})\prime (x_i - \hat{\boldsymbol{\mu}}_{SD})}{\sum_{1=1}^{n} \omega_i}$$

donde $\omega_i(r_i)\colon (0, +\infty) \rightarrow \ (0, +\infty)$es una función de peso que penaliza o reduce el peso de las observaciones con una gran atipicidad. Hay varios enfoques sobre la elección de las ω_i para la forma de estas funciones. Una familia de funciones de peso utilizadas en la literatura son los "pesos de Huber" que mejoran el rendimiento en la detección de valores atípicos

El Estimador SD es especialmente útil en el contexto de datos asimétricos y/o con presencia de outliers, ya que se ajusta a las características de la distribución y proporciona estimaciones robustas de la tendencia central. Al ser menos influenciado por los valores atípicos y las distribuciones no normales, el Estimador SD ofrece una alternativa más fiable y precisa en comparación con los estimadores tradicionales.

En resumen, el Estimador Stahel-Donoho (SD) es un estimador robusto utilizado para calcular medidas de tendencia central en presencia de datos asimétricos y/o contaminados por outliers. Se adaptar a diferentes proporciones de datos contaminados y niveles de asimetría. El Estimador SD proporciona estimaciones más confiables y resistentes a los outliers que los estimadores convencionales.

5.2.2. Metodología Outlyingness Skewness Adjusted (SAO)

La metodología Outlyingness Skewness Adjusted (SAO), propuesta por Mia Hubert y Stephan Van der Veeken en 2008, es una técnica robusta de detección de outliers que se enfoca en abordar la asimetría en los datos. Esta metodología combina la detección de outliers basada en la desviación mediana absoluta (MAD) con un ajuste de la asimetría de los datos. La SAO se basa en el principio de que la asimetría puede afectar la detección de outliers, ya que los valores extremos en una cola larga de la distribución pueden no ser correctamente identificados utilizando técnicas convencionales. Los autores propusieron el SAO como un método alternativo de búsqueda de proyección para la detección de valores atípicos en distribuciones sesgadas.

El SAO se define como:

$$SAO(\,x_i, \mathbf{X}) = \sup_{r \in S_p} \begin{cases} \dfrac{r'x_i - med(r'X)}{\omega_2 - med(r'X)} & r'x_i > med(r'X) \\ \dfrac{med(r'X)\ -\ r'x_i}{med(r'X) - \omega_1} & r'x_i < med(r'X) \end{cases}$$

donde ω_1 y ω_2 denotan el bigote inferior y superior del diagrama de caja ajustado (Hubert y Vandervieren, 2008), respectivamente.

Si $\mathrm{SAO}(\mathrm{x}_i) > Q_3(\mathrm{SAO}) + 1.5e^{3MC}\, IQR(\mathrm{SAO})$ el individuo x_i es etiquetado como un valor atípico. Aquí MC denota la estadística de MedCouple (Brys et al., 2004).

La Metodología SAO proporciona una forma robusta de detectar outliers al tener en cuenta tanto la desviación mediana absoluta como el ajuste de la asimetría de los datos. Al considerar la asimetría, la SAO es capaz de detectar outliers en distribuciones asimétricas y proporcionar resultados más confiables en presencia de datos no normales. Esta metodología es especialmente útil en situaciones donde la asimetría es una característica relevante de los datos y puede afectar la detección tradicional de outliers.

5.2.3. Estadístico Robusto de Asimetría MedCouple

El estadístico robusto de asimetría, conocido como MedCouple, es una medida robusta utilizada para evaluar la asimetría en una distribución de datos. A diferencia de los métodos tradicionales como el coeficiente de asimetría de Pearson, el MedCouple es menos sensible a los outliers y proporciona una estimación más confiable de la asimetría en datos que contienen valores extremos o distribuciones no simétricas.

El MedCouple se basa en la mediana y se utiliza para caracterizar la asimetría en términos de la dirección y la magnitud de la cola de la distribución. El valor del MedCouple puede variar entre -1 y 1, donde:

- Un MedCouple de 0 indica una distribución simétrica.
- Un MedCouple positivo indica una asimetría hacia la derecha, lo que implica que la cola de la distribución se extiende más hacia valores mayores que la mediana.
- Un MedCouple negativo indica una asimetría hacia la izquierda, lo que implica que la cola de la distribución se extiende más hacia valores menores que la mediana.

El cálculo del MedCouple implica varios pasos:

- Ordenar los datos de forma ascendente.
- Calcular las distancias desde cada dato hasta la mediana.
- Comparar estas distancias para determinar la asimetría de la distribución.

El MedCouple se calcula mediante la siguiente fórmula:

$$\text{MedCouple} = \frac{(Q_3 - mediana) - (mediana - Q_1)}{Q_3 - Q_1}$$

donde Q_1 es el primer cuartil (25%) y Q_3 es el tercer cuartil (75%) de los datos.

El MedCouple puede utilizarse para identificar la presencia y la dirección de la asimetría en una distribución. Un valor cercano a 0 indica una distribución simétrica, mientras que valores más alejados de 0 indican una asimetría más pronunciada. Además, el MedCouple es resistente a los outliers, lo que significa que su cálculo no se ve afectado significativamente por valores atípicos o extremos.

En resumen, el Estadístico robusto de asimetría (MedCouple) es una medida robusta utilizada para evaluar la asimetría en los datos. Proporciona una estimación más confiable de la asimetría en comparación con los métodos tradicionales y es menos sensible a los outliers. El MedCouple es útil en estudios donde la asimetría de los datos es una característica importante para tener en cuenta, como en la detección de outliers con técnicas robustas.

5.2.4. Boxplot para Datos Asimétricos

Un boxplot ó diagrama de caja y bigotes, es una representación gráfica que proporciona información sobre la distribución y los valores atípicos de un conjunto de datos. Es especialmente útil para identificar características de asimetría en los datos, como colas largas y sesgos. Para datos asimétricos, el boxplot presenta una forma característica que refleja la asimetría de la distribución.

La caja en el centro del boxplot representa el rango intercuartílico (IQR, por sus siglas en inglés), que es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1) de los datos. La longitud de la caja indica la dispersión de los valores en el rango intercuartílico. Si la caja está desplazada hacia la parte inferior del gráfico, indica una asimetría positiva (cola hacia la derecha), mientras que, si está desplazada hacia la parte superior, indica una asimetría negativa (cola hacia la izquierda). Los bigotes, que se extienden desde la caja, representan los valores mínimos y máximos dentro de un rango específico. Por lo general, se definen como 1,5 veces el rango intercuartílico.

$$\text{IQR} = Q_3 - Q_1$$

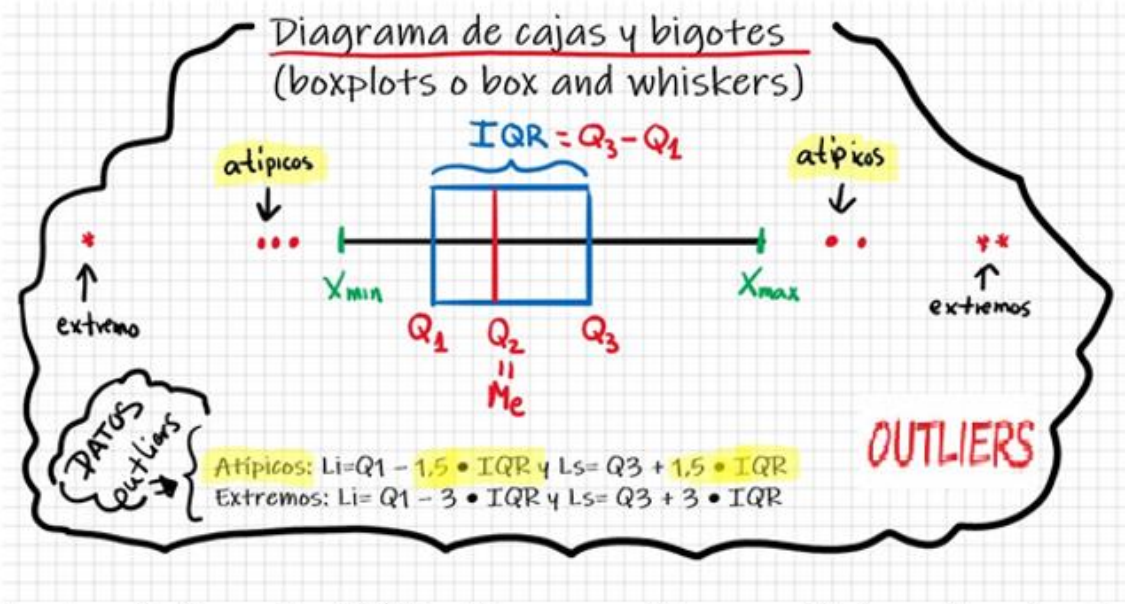
Los límites de los bigotes se calculan como:

$$\text{Límite Inferior} = Q_1 - 1.5 * IQR$$

$$\text{Límite Superior} = Q_3 + 1.5 * IQR$$

Sin embargo, en el caso de datos asimétricos, los bigotes pueden ser más cortos en un lado (si hay una cola corta) y más largos en el otro (si hay una cola larga). Los valores que están más allá de los bigotes se consideran valores atípicos o outliers y se representan como puntos o asteriscos en el gráfico. Los outliers pueden ser indicativos de la presencia de datos extremos o inusuales en la distribución.

Figura No. 3. Gráfico diagrama de caja o boxplot.



Fuente: Benítez, E. (2021). *Diagramas Cajas y Bigotes.* Tomado de https://matesnoaburridas.wordpress.com/2021/03/28/diagramas-cajas-y-bigotes/

En resumen, un boxplot para datos asimétricos mostrará una caja desplazada hacia un lado y bigotes desiguales debido a la asimetría en la distribución. Este tipo de gráfico proporciona una visualización útil para comprender y detectar la presencia de asimetría en los datos, así como la identificación de posibles outliers.

6. Estado del Arte.

Juan y Prieto (1995) presentan un innovador método de submuestreo para el cálculo de estimadores multivariados con un alto punto de quiebre (breakdown point), centrado en la robustez de los estimadores estadísticos ante la presencia de datos contaminados o valores atípicos. La metodología implica la generación de submuestras aleatorias y la aplicación de un algoritmo iterativo que permite obtener estimaciones robustas y eficientes. Cada submuestra se analiza de manera independiente, y las estimaciones resultantes se combinan para producir un estimador global resistente a la influencia de outliers. Este enfoque asegura que las estimaciones finales sean menos susceptibles a la distorsión causada por datos extremos, mejorando la precisión y fiabilidad de los resultados en análisis multivariados. El método de submuestreo propuesto es una contribución significativa al campo de la estadística robusta, proporcionando una herramienta efectiva para obtener resultados precisos y representativos, incluso en entornos contaminados, lo que es esencial en aplicaciones prácticas donde la integridad de los datos puede ser comprometida.

Maronna y Yohai (1995) presentan un detallado estudio sobre el comportamiento del estimador robusto multivariado propuesto por Stahel-Donoho, un método innovador diseñado para mejorar la detección y estimación de valores atípicos en datos multivariados. El estudio analiza las propiedades del estimador en términos de robustez, eficiencia y consistencia, evaluando su desempeño bajo diferentes escenarios de datos. Una de las principales ventajas del estimador Stahel-Donoho es su alto punto de quiebre (breakdown point), lo que significa que puede manejar un gran porcentaje de datos contaminados sin perder su capacidad para producir estimaciones precisas. Los resultados del estudio demuestran que el estimador Stahel-Donoho no solo es robusto frente a la presencia de valores atípicos, sino que también mantiene una alta eficiencia y consistencia en la estimación de parámetros estadísticos. Esta robustez y eficiencia lo hacen particularmente útil en aplicaciones prácticas donde los datos pueden estar contaminados o ser inherentemente asimétricos. El análisis de Maronna y Yohai subraya la eficacia del estimador Stahel-Donoho, proporcionando evidencia de su capacidad para mejorar la fiabilidad de los análisis estadísticos en contextos multivariados.

Brys, Hubert y Struyf (2004) presentan una innovadora medida robusta de asimetría diseñada específicamente para datos sesgados. A diferencia de las medidas tradicionales de asimetría, que pueden ser altamente sensibles a valores atípicos, la medida propuesta por estos autores se basa en una combinación de la mediana absoluta de las desviaciones y la mediana absoluta de las desviaciones en el sentido contrario. Este enfoque dual permite que la medida sea resistente a la influencia de outliers, proporcionando una estimación más precisa y confiable de la asimetría inherente en los datos. En sus estudios, Brys, Hubert y Struyf demuestran que esta nueva medida robusta de asimetría supera significativamente a las medidas tradicionales en términos de estabilidad y precisión, especialmente en conjuntos de datos con una distribución no simétrica y con la presencia de valores atípicos. La investigación concluye que la implementación de esta medida puede mejorar considerablemente la calidad de los análisis estadísticos y la interpretación de datos sesgados, ofreciendo una herramienta valiosa para estadísticos y analistas de datos que trabajan con distribuciones complejas.

Peña y Prieto (2007) proponen un enfoque innovador para la detección de valores atípicos y la estimación robusta en conjuntos de datos multivariados de alta dimensionalidad. Este método combina el uso de direcciones aleatorias y específicas, lo que permite una

identificación más precisa de outliers y una estimación más robusta de los parámetros. La metodología se basa en la generación de múltiples direcciones de proyección, lo que facilita la captura de estructuras complejas en los datos que podrían pasar desapercibidas con enfoques unidimensionales. Posteriormente, se calculan estadísticas robustas para cada proyección y se comparan con umbrales predefinidos, lo que permite determinar la presencia de valores atípicos. La combinación de estas técnicas no solo mejora la precisión en la detección de outliers, sino que también optimiza la robustez de las estimaciones en presencia de datos contaminados o distribuciones no normales. Los resultados de su estudio indican que este enfoque es altamente efectivo para manejar la alta dimensionalidad de los datos, proporcionando una herramienta poderosa para el análisis multivariado en diversos campos de investigación.

Hubert y Vandervieren (2008) proponen un método innovador para ajustar el boxplot tradicional a fin de manejar distribuciones sesgadas de manera más efectiva. Este boxplot ajustado considera explícitamente la asimetría de los datos, proporcionando así una visualización más precisa y representativa de la distribución subyacente. La técnica consiste en transformar los datos utilizando una corrección que permite calcular límites más adecuados para el boxplot. Esta transformación ajusta las colas de la distribución, lo que mejora la identificación de valores atípicos y ofrece una mejor comprensión de la estructura de los datos sesgados. El boxplot ajustado se muestra particularmente útil en contextos donde las distribuciones no son simétricas, ya que proporciona una herramienta robusta para visualizar y analizar datos de manera efectiva. Esta metodología ha demostrado ser valiosa para investigadores y analistas que trabajan con datos sesgados, permitiendo una detección más precisa de outliers y una interpretación más fiel de las características de la distribución.

Hubert y Van der Veeken (2008) proponen un enfoque innovador para abordar la detección de valores atípicos en conjuntos de datos con distribuciones asimétricas. Reconociendo las limitaciones de las técnicas tradicionales de detección de outliers en presencia de asimetría significativa, los autores desarrollan el método "Skewness Adjusted Outlyingness" (SAO). Este método combina la robusta medición de la mediana absoluta de las desviaciones (MAD) con un ajuste específico para corregir la asimetría de los datos. A través de estudios experimentales detallados y comparaciones con otros métodos existentes, Hubert y Van der Veeken demuestran la efectividad y las ventajas de SAO en la detección precisa de valores atípicos en datos asimétricos. Este enfoque no solo mejora la sensibilidad para identificar outliers en distribuciones no normales, sino que también ofrece una herramienta valiosa para la minería de datos y la estadística aplicada, contribuyendo significativamente al campo de la detección robusta de outliers.

Van Aelst, Vandervieren y Willems (2011) proponen una extensión del estimador Stahel-Donoho mediante la incorporación de pesos específicos para cada observación o celda de datos. Estos pesos se determinan considerando la contribución de cada observación a la estructura de la distribución de datos. La metodología desarrollada tiene como objetivo mejorar las estimaciones robustas en presencia de diferentes tipos de contaminación o asimetría en los datos. Los autores presentan ejemplos y estudios de simulación que ilustran la utilidad y el rendimiento de los estimadores Stahel-Donoho con pesos celulares. Este enfoque demuestra la capacidad del estimador para adaptarse a diversos escenarios de datos, asegurando estimaciones más precisas y confiables bajo condiciones no estándar.

Van Aelst, Vandervieren y Willems (2012) introducen un nuevo estimador basado en la metodología Stahel-Donoho, el cual emplea el concepto de "outlyingness" (grado de ser un valor atípico) junto con una función de influencia robusta de tipo Huber. Este enfoque innovador combina la robustez inherente a la metodología Stahel-Donoho con una medida de la presencia de outliers que se adapta de manera más efectiva a diversas estructuras de datos. El estimador propuesto se caracteriza por su capacidad para proporcionar estimaciones precisas y robustas incluso en presencia de datos contaminados o con asimetría, aspectos críticos en el análisis de datos no convencionales. Los autores detallan exhaustivamente el procedimiento de cálculo del estimador, destacando su proceso de minimización de la función de riesgo influenciada por los outliers a través de la función de influencia robusta de Huber. Además, presentan resultados empíricos que evidencian la eficacia y superioridad del estimador propuesto frente a otros métodos existentes en términos de precisión y robustez en diferentes escenarios de datos.

Van Aelst (2016) aborda la estimación Stahel-Donoho en el contexto de datos de alta dimensionalidad. Se discute cómo la presencia de muchas variables puede afectar la identificación y estimación de outliers, así como la elección de la medida de distancia adecuada para medir la proximidad entre observaciones. Se presenta una adaptación del estimador Stahel-Donoho para el caso de datos de alta dimensión, junto con resultados teóricos que respaldan su rendimiento.

Tabla No. 1. Resumen de los criterios de comparación entre los artículos seleccionados.

Articulos	Año de Publicació	Familia de Tecnicas	Estimadores	Datos	Conclusión
Proyecto de Grado	2024	Técnica basadas en el estimador Stahel-Donoho.	Estimadores Robustos	Simulaciones	
A Subsampling Method for the Computation of Multivariate Estimators with High Breakdown Point	1995	Remuestreo y Submuestreo	Estimadores Robustos	Simulaciones y análisis empiricos	El método de submuestreo propuesto proporciona resultados precisos y fiables, incluso en presencia de datos contaminados. Los resultados obtenidos con el método son comparables con otros estimadores robustos de alto punto de quiebre pero con menores requerimientos computacionales.
The Behavior of the Stahel-Donoho Robust Multivariate Estimator	1995	Análisis teóricos y estudios empíricos.	Estimadores Robustos	Simulaciones y análisis empiricos	El estimador Stahel-Donoho es un estimador multivariante robusto y eficiente, adecuado para situaciones con datos contaminados y múltiples variables anómalas, destacándose por su alto punto de quiebre.
A Robust Measure of Skewness	2004	Métodos y medidas robustas para el análisis de datos.	Estimadores Robustos	Simulaciones y análisis empiricos	La medida robusta de asimetría desarrollada por Brys, Hubert y Struyf es una herramienta poderosa para evaluar la asimetría de una distribución de datos de manera precisa y fiable, especialmente en presencia de datos atípicos. Esta medida es menos sensible a los valores extremos y más representativa de la estructura central de los datos.
Combining Random and Specific Directions for Outlier Detection and Robust Estimation in High-Dimensional Multivariate Data	2007	Técnicas para la detección de outliers y la estimación robusta en datos multivariados de alta dimensionalidad.	Estimadores Robustos	Simulaciones y datos reales	El método de combinar direcciones aleatorias y específicas propuesto por Peña y Prieto es una herramienta robusta y eficaz para la detección de valores atípicos y la estimación en datos multivariantes de alta dimensión. El enfoque mejora significativamente la capacidad de detectar valores atípicos y proporciona estimaciones robustas, todo ello manteniendo una eficiencia computacional
An adjusted boxplot for skewed distributions	2008	Métodos de visualización y detección de outliers en datos sesgados.	Estimadores Robustos	Simulaciones y análisis empiricos	El boxplot ajustado propuesto por Hubert y Vandervieren es una mejora significativa sobre el boxplot tradicional para distribuciones sesgadas. Este método proporciona una detección más precisa de valores atípicos y una representación gráfica más fiel de la dispersión de datos en distribuciones asimétricas, ampliando así la utilidad del boxplot en análisis estadísticos exploratorios.
Outlier detection for skewed data	2008	Métodos de detección de outliers específicamente diseñados para datos sesgados.	Estimadores Robustos	Simulaciones y análisis empiricos	El método propuesto por Hubert y Van der Veeken es una herramienta efectiva y robusta para la detección de valores atípicos en conjuntos de datos sesgados. Este enfoque utiliza estimaciones de densidad de probabilidad robustas y medidas de distancia para identificar valores atípicos de manera confiable, incluso en distribuciones altamente asimétricas.
A Stahel–Donoho estimator based on huberized outlyingness	2012	Técnica basada en el estimador Stahel-Donoho que utiliza la medida de "outlyingness" con una modificación de Huberización.	Estimadores Robustos	Simulaciones y análisis empiricos	El estimador robusto basado en la huberización de la medida de excepcionalidad propuesto por Hubert y Van der Veeken es una mejora significativa sobre el estimador Stahel-Donoho original. Este tiene una mayor resistencia a la presencia de valores atípicos y ofrece una opción confiable para la estimación robusta en conjuntos de datos multivariados contaminados.
Stahel-Donoho estimators with cellwise weights	2012	Métodos de estimación robusta que incorporan pesos a nivel de celda.	Estimadores Robustos	Simulaciones	El estimador Stahel-Donoho con pesos celulares propuesto por Van Aelst, Vandervieren y Willems ofrece una mejora significativa sobre el estimador original al permitir una adaptación más flexible a la estructura de los datos y la presencia de valores atípicos.
Stahel–Donoho estimation for high-dimensional data	2016	Técnicas de estimación basadas en el enfoque Stahel-Donoho para datos de alta dimensionalidad.	Estimadores Robustos	Simulaciones y análisis empiricos	El estimador Stahel–Donoho puede ser una herramienta que proporciona estimaciones robustas y precisas incluso en situaciones complejas. Se requieren adaptaciones y extensiones específicas para abordar los desafíos asociados con la alta dimensionalidad de los datos, lo que abre oportunidades para futuras investigaciones en este campo.

Fuente: Elaboración propia.

7. Marco Metodológico.

7.1. Comprensión de la Investigación.

En el contexto del análisis de datos multivariantes y de alta dimensionalidad, es esencial contar con herramientas estadísticas robustas capaces de manejar la presencia de valores atípicos y distribuciones sesgadas. En la literatura reciente, se han propuesto y validado varios métodos destinados a mejorar la robustez y precisión de los estimadores en estos escenarios desafiantes. Entre los enfoques destacados se incluyen técnicas de subsampling, como las propuestas por Juan y Prieto (1995), que permiten la computación eficiente de estimadores robustos. Asimismo, Maronna y Yohai (1995) han investigado la generación de datos con diferentes niveles de contaminación para evaluar la resistencia de los estimadores. Brys, Hubert y Struyf (2004) han empleado medidas basadas en cuantiles para mitigar la influencia de valores atípicos, mientras que Peña y Prieto (2007) han combinado direcciones aleatorias y específicas para la detección de outliers.

En paralelo, se han desarrollado técnicas como los boxplots ajustados propuestos por Hubert y Vandervieren (2008), que adaptan los límites del boxplot para reflejar distribuciones asimétricas. Además, Hubert y Van der Veeken (2008) han utilizado estimaciones robustas de densidad y medidas de distancia para identificar valores atípicos en distribuciones asimétricas. Van Aelst, Vandervieren y Willems (2012) han introducido pesos celulares para adaptar el estimador Stahel-Donoho a la estructura específica de los datos, mientras que Van Aelst (2016) ha explorado técnicas de regularización y reducción de dimensionalidad para datos de alta dimensión. Estos estudios no solo han demostrado su eficacia mediante simulaciones y aplicaciones empíricas, sino que también ofrecen soluciones prácticas para diversos desafíos relacionados con datos contaminados.

7.2. Comprensión de las Técnicas.

El método Stahel-Donoho (SD), propuesto por Stahel (1981) y Donoho (1982), es una técnica avanzada utilizada para la detección de valores atípicos en datos multivariados y para realizar estimaciones robustas. Este método implica la generación aleatoria de un número fijo de direcciones y la evaluación de la medida de periferia para los datos proyectados en cada una de estas direcciones. Una característica sobresaliente del algoritmo SD es su invarianza afín, lo que asegura que sus resultados no se vean afectados por transformaciones lineales de los datos. Además, presenta un alto punto de quiebre, lo que proporciona una robustez considerable frente a una alta proporción de valores atípicos, como se observó por Maronna y Yohai (1995), y eficiencia asintótica, tal como se describe en términos de su función de influencia por Gervini (2002).

A pesar de sus ventajas, el método Stahel-Donoho presenta limitaciones significativas, como la alta carga computacional requerida debido al cálculo intensivo de un gran número de direcciones, lo cual puede afectar su eficacia en escenarios con datos de alta dimensionalidad (p). En respuesta a estas limitaciones, se ha desarrollado el método de asimetría ajustada outlyingness (SAO), propuesto por Hubert y Van der Veeken (2008), que capitaliza la asimetría mediante el ajuste de la medida de periferia, permitiendo así una adaptación precisa a la estructura asimétrica de los datos. Esta capacidad de ajuste hace que SAO sea más sensible a los valores atípicos en distribuciones sesgadas y menos reactivo a las desviaciones en distribuciones simétricas, proporcionando una herramienta robusta y eficaz para la detección de valores atípicos en contextos donde la asimetría juega un papel crucial.

7.3. Preparación de los Datos.

Los estudios referenciados han utilizado diversas estrategias para preparar los datos y mejorar la robustez de los estimadores propuestos. Estas incluyen técnicas avanzadas como el uso de submuestreo, generación de datos contaminados, métodos basados en cuantiles, combinación de direcciones aleatorias y específicas, ajuste de límites de boxplot, estimaciones robustas de densidad, medidas de distancia adaptativas, introducción de pesos celulares y técnicas de regularización y reducción de dimensionalidad. Estas metodologías han sido aplicadas con el objetivo de asegurar la precisión y fiabilidad de los estimadores en presencia de valores atípicos y estructuras complejas de datos multivariados y de alta dimensionalidad.

7.4. Metodología

7.4.1. La Asimetría como Índice de Proyección.

El coeficiente de asimetría es otro índice estadístico relevante para la detección de outliers multivariados mediante la búsqueda de proyecciones. Recientemente, han surgido enfoques que emplean la asimetría muestral. Loperfido (2018) propuso un procedimiento para calcular las direcciones que maximizan la asimetría muestral como una herramienta para el análisis exploratorio de datos, incluyendo la detección tentativa de outliers. Además, Ortiz (2019) desarrolló un método de detección de outliers multivariados basado en la proyección que maximiza el tercer momento absoluto muestral de los datos proyectados, siguiendo una iteración matricial basada en eigenvectores como lo propusieron Peña y Prieto (2001).

Sea $\boldsymbol{z} = \left(z_1, z_2, \dots, z_n\right) \in \mathbb{R}^{pxn}$ una muestra contaminada de tamaño n en un espacio p-dimensional. Se asume que esta muestra es una mezcla de dos distribuciones de probabilidad:$(1 - \alpha) + \alpha\mathcal{G}, \mathcal{F}$ donde α representa la proporción de contaminación, $\mathcal{F}$ es la distribución de probabilidad de los puntos que no son valores atípicos, y $\mathcal{G}$ es la distribución de probabilidad de los valores atípicos. Además, se asume que la muestra $\boldsymbol{z}$ ha sido centrada y escalada. Para complementar, se hace uso de $m_3(\cdot)$ para representar el coeficiente del tercer momento muestral univariado.

Ahora, consideremos $\boldsymbol{r} \in \mathbb{R}^{px1}$ como un vector p-dimensional desconocido, y sea $\boldsymbol{d_1}$ la dirección obtenida al resolver el siguiente problema de optimización:

$$\boldsymbol{r_s} = \underset{\boldsymbol{r}}{\operatorname{argmax}} \left(\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{r}' z_i)^3 \right)^2, \qquad s.t. \;\; \boldsymbol{r}'\boldsymbol{r} = 1 \tag{1}$$

Cabe destacar que es mucho mejor maximizar $m_3(\cdot)^2$ en lugar de $m_3(\cdot)$ o $|m_3(\cdot)|$, ya que en el caso de una contaminación asimétrica $\boldsymbol{r_s}$ se enfrenta a una asimetría positiva o negativa extrema en comparación con $m_3(\cdot)$. Por otro lado, maximizar $m_3(\cdot)^2$ es más sencillo que $|m_3(\cdot)|$. El procedimiento de maximización en la Ecuación (1) se puede calcular mediante un procedimiento de Newton-Raphson o una iteración de matriz de valores propios; esta última de acuerdo con las condiciones de primer orden y la función Lagrangiana del problema de optimización. En este caso, se sigue la opción de iteración de matriz de valores propios, como propuso Ortiz (2019).

7.4.2. r_s como Generador de Direcciones Aleatorias.

Peña y Prieto (2007) afirman que una dirección generada por dos observaciones, en un modelo de contaminación asimétrica, una del grupo de puntos no atípicos y la otra del grupo de puntos atípicos, podría ser una dirección aleatoria inicial interesante. Dado que las proyecciones en esta dirección tienden a ordenar los puntos, ubicando la contaminación en un lado de la proyección, ya sea valores atípicos a la izquierda o a la derecha.

Un aspecto muy importante es cuántas direcciones aleatorias son necesarias o suficientes para mejorar la detección de valores atípicos. Varios autores han discutido sobre el número de direcciones en algunas otras metodologías estadísticas. Hubert y Van der Veeken (2008) argumentaron que 250*p* direcciones aleatorias son computacionalmente adecuadas para su propuesta de Atipicidad Ajustada de Outliers. Por lo tanto, se propone una versión modificada del procedimiento de Atipicidad Ajustada de Outliers, de la siguiente manera:

Sea $\boldsymbol{y} = \mathbf{r}'\mathbf{z}$ la muestra multivariada proyectada en la dirección que maximiza la asimetría al cuadrado $\boldsymbol{r_s}$. Por lo tanto, $\boldsymbol{y} = (y_1, y_2, \dots, y_n) = (\boldsymbol{r_s}'z_1, \boldsymbol{r_s}'z_2, \dots, \boldsymbol{r_s}'z_n) \in \mathbb{R}^{1xn}$.

1. Dibujar una muestra aleatoria (A_1) de tamaño $2p$ tomada del conjunto $\{Y_{[1]}, \dots, Y_{[n/4]}\}$.
2. Dibujar una muestra aleatoria (A_2) de tamaño $2p$ tomada del conjunto $\{Y_{[3n/4]}, \dots, Y_{[n]}\}$.
3. Tomar las observaciones multivariadas de $\boldsymbol{z}$ que corresponden a los individuos seleccionados de los conjuntos A_1 y A_2.
4. Generar las direcciones unitarias a partir de todos los pares posibles de puntos multivariados, uno de A_1y otro de A_2, es decir, $\boldsymbol{u}_k = \frac{z_i(A_1) - z_j(A_2)}{\|z_i(A_1) - z_j(A_2)\|}$ para $k = 1, \dots, 4p^2$. Se obtiene $4p^2$ direcciones de proyección.
5. Finalmente, denote el conjunto $S_p^* = \{\boldsymbol{r_s}, \boldsymbol{u}_1 \dots, \boldsymbol{u}_{4p^2}\}$, el conjunto de direcciones para el cálculo de la medida atípica (SAO).

7.4.3. Direcciones Aleatorias Específicas SAO (RSP-SAO).

Una vez que se construye el conjunto S_p^*, se propone el siguiente procedimiento, denominado RSP-SAO, para la detección de valores atípicos multivariados:

1. Calcular la medida de atipicidad SDO* utilizando el estadístico Q_n (Rousseeuw y Croux, 1993), en lugar de la MAD.

$$\mathrm{SDO}^*(\mathrm{z}_i, \mathbf{z}) = \max_{a \in S_p^*} \frac{|\mathbf{a}'\mathrm{z}_i - med(\mathbf{a}'\mathbf{z})|}{Q_n(\mathbf{a}'\mathbf{z})}$$

2. Tomar las medidas centrales o más profundas de atipicidad como $\mathrm{SDO}_d = \{\mathrm{SDO}^*_{[1]}, \dots, \mathrm{SDO}^*_{[3n/5]}\}$.
3. Si $\mathrm{SDO}^*(\mathrm{x}_i) > Q_3(\mathrm{SDO}_d) + 1.5e^{3MC(\mathrm{SDO}_d)}\, IQR(\mathrm{SDO}_d)$ el individuo z_i es etiquetado como un valor atípico. El lado derecho de la desigualdad se refiere al bigote superior del boxplot ajustado (Hubert y Vandervieren, 2008), donde MC denota el estadístico medcouple, un estimador robusto de la asimetría (Brys et al., 2004).

7.5. Modelado

En este caso por ser un proyecto de investigación en contraste de técnicas se seleccionan y aplican técnicas robustas de detección de outliers, como el uso de estadísticas robustas o métodos basados en medianas. También se generan escenarios simulados para evaluar el desempeño de las técnicas seleccionadas según las características de los datos y los objetivos del estudio. Finalmente se aplican las técnicas de detección de outliers en los datos preparados utilizando las técnicas robustas elegidas.

7.5.1. Función Sk.Child

La primera función Sk.Child (ver Apéndice A) se utiliza para calcular direcciones aleatorias y específicas basadas en la dirección de máxima asimetría de un conjunto de datos multivariado, utilizando el método de proyección máxima (Ortiz y Londoño, 2023).

Las entradas necesarias para la función son una matriz de datos estandarizados y una proyección de máxima asimetría. La función produce como salida un conjunto de direcciones aleatorias basadas en la asimetría de los datos.

Inicialmente, la función determina el número de variables en la matriz de entrada y calcula los percentiles 25 y 75 de la proyección. A continuación, selecciona muestras aleatorias de las posiciones donde la proyección es menor o igual al percentil 25 y mayor o igual al percentil 75, generando subconjuntos de la matriz de entrada correspondientes a estas posiciones.

Para calcular las direcciones, la función crea una matriz de diferencias entre los subconjuntos seleccionados, normaliza cada fila para obtener direcciones unitarias, y las agrega a una lista de direcciones. Finalmente, la función retorna la transpuesta de esta lista, proporcionando así las direcciones aleatorias basadas en la asimetría de los datos.

7.5.2. Función adj.outly

La segunda función adj.outly (ver Apéndice B) tiene la finalidad de calcular una medida de "outlyingness" ajustada para un conjunto de datos univariados (Ortiz y Londoño, 2023).

Esta función toma como entrada un vector X de tamaño n que contiene los datos y devuelve un vector Adj.AO del mismo tamaño, que contiene la medida de outlyingness para cada observación en X. El proceso de la función se desarrolla en varios pasos, comenzando con el cálculo de parámetros robustos y terminando con el retorno del vector ajustado.

En los primeros pasos, la función calcula parámetros robustos esenciales: la mediana de X (mu.rob), el Medcouple (mc), que es una medida de asimetría robusta, el rango intercuartílico (iqr), y la longitud de X(n). Luego, se inicializa un vector de ceros Adj.AO de tamaño n. Posteriormente, se calculan los límites ajustados (w_1 y w_2). Si mc es mayor o igual a 0, los límites se calculan usando el primer y tercer cuartil de X, ajustados por factores basados en mc y iqr. Si mc es menor que 0, los límites se calculan de manera similar, pero con factores diferentes.

Finalmente, la función calcula la medida de outlyingness ajustada para cada elemento en X. Para cada elemento i, si X_i es mayor que mu.rob, la outlyingness se calcula como (X_i -

mu.rob) / (w_2 - mu.rob). Si X_i es menor o igual a mu.rob, la outlyingness se calcula como (mu.rob - X_i) / (mu.rob - w_1).

El resultado es el vector Adj.AO, que contiene la medida de outlyingness ajustada para cada observación en X, y este vector se devuelve como la salida de la función.

7.5.3. Función Adj.Outlier

La tercera función Adj.Outlier (ver Apéndice C) que detecta valores atípicos univariados basándose en la asimetría ajustada de un conjunto de datos univariado (Ortiz y Londoño, 2023).

La función toma como entrada un vector de datos univariados (X_1) y un tipo de calibración que puede ser "1" para usar Medcouple o cualquier otro valor para usar el método tradicional del "upper whisker". La salida es un vector binario lab.out donde "1" indica un valor atípico y "0" indica un valor no atípico.

El proceso comienza filtrando los datos para incluir solo aquellos valores que están por debajo o igual al percentil 60 de X_1. Luego, se inicializan varios parámetros: la longitud del vector de datos, un vector de ceros para las etiquetas de los outliers, la medida de asimetría robusta (Medcouple), el rango intercuartílico (IQR) y el percentil 75 de los datos filtrados. Dependiendo del tipo de calibración especificado, se calcula un punto de corte (cutoff) usando fórmulas diferentes. El punto de corte incluye un ajuste basado en la asimetría; de lo contrario, se utiliza la fórmula tradicional.

Finalmente, la función identifica los valores atípicos como aquellos valores en X_1 que son mayores o iguales al punto de corte. Las posiciones correspondientes en el vector de etiquetas se actualizan para marcar estos valores como atípicos. La función devuelve el vector de etiquetas binarias (lab.out), proporcionando una forma clara y precisa de identificar outliers en un conjunto de datos univariado.

7.5.4. Función max_skew

La cuarta función max_skew (ver Apéndice D) diseñada para calcular la proyección de máxima asimetría en un conjunto de datos multivariado (Ortiz y Londoño, 2023).

La función toma como entrada una matriz X de tamaño n*p, con observaciones por filas y variables por columnas. Su salida consiste en un vector dir_vec que maximiza el coeficiente de asimetría de los datos en X y el valor skew_val, que representa el tercer momento central estandarizado de la proyección de X en dir_vec.

La función comienza con una verificación inicial para asegurarse de que el argumento X esté definido, seguida de la definición de una tolerancia tol para los cálculos y la obtención de las dimensiones de X (n.x y p.x). Luego, los datos son estandarizados y transpuestos. Se normalizan las columnas de datos, se calculan los componentes principales y se selecciona el vector propio con el mayor momento cúbico central absoluto como vector de proyección inicial (d0). El algoritmo de búsqueda se inicia con la definición de la matriz de covarianza y su raíz cuadrada inversa, iterando hasta que se cumpla la tolerancia o se alcance un máximo de 200 iteraciones, actualizando el vector de proyección d0 basado en los vectores propios de la matriz de asimetría.

Finalmente, la función selecciona la mejor dirección de proyección comparando las proyecciones calculadas y seleccionando la mejor dirección (best_of_all). Se calcula el tercer momento central estandarizado de la proyección de X en dir_vec, retornando una lista que contiene dir_vec, el vector de proyección que maximiza la asimetría, y skew_val, el valor del tercer momento central estandarizado de dicha proyección.

7.5.5. Función val_skew

La quinta función val_skew (ver Apéndice E) cuya finalidad es evaluar el coeficiente de momento de orden k para la proyección univariada de datos multivariados (Peña y Prieto, 2001).

La función requiere tres entradas: una matriz x con observaciones por filas, un vector d que representa la dirección de proyección, y el orden del momento km, con un valor predeterminado de 3. La salida es el valor del coeficiente de momento (mc) calculado para los datos univariados proyectados.

En su ejecución, la función realiza varios pasos clave. Primero, inicializa y verifica las dimensiones de la matriz x y el vector d para asegurar la compatibilidad entre las dimensiones de las observaciones y las direcciones de proyección. Si las dimensiones no coinciden, la función lanza un error. Luego, proyecta los datos multivariados en la dirección especificada por d y calcula la media y la desviación absoluta de la proyección resultante respecto a la media.

Finalmente, la función calcula la varianza y el momento de orden km de la proyección. Utiliza estos valores para obtener el coeficiente de momento ajustado, que se calcula como el momento de orden km dividido por la varianza elevada a la potencia (km/2). El coeficiente de momento ajustado (mc) proporciona una medida de asimetría u otros momentos de los datos proyectados, y es devuelto como la salida de la función.

En resumen, val_skew proyecta los datos multivariados en una dirección dada y calcula el coeficiente de momento para la proyección resultante, ofreciendo una medida útil para el análisis de asimetría y otros momentos estadísticos.

7.5.6. Función gen_rcorr

La sexta función gen_rcorr (ver Apéndice F) tiene el objetivo de generar una matriz de correlación aleatoria para una variable aleatoria multivariada de dimensión p (Velasco, 2023).

La función toma dos entradas: cond.S, que representa la condición de la matriz de correlación deseada, y p.x, que es la dimensión de la variable aleatoria multivariada (número de variables). La salida es una matriz de correlación Sigma que refleja la estructura de correlación aleatoria generada.

La función comienza con una serie de pasos de inicialización: define el número máximo de iteraciones (maxits = 100) y una tolerancia (tol = 1.0e-5), verifica que p.x sea mayor que 2 y lanza un error si no se cumple esta condición. Luego, genera los valores propios (lambda) y una matriz aleatoria x de tamaño p.x x p.x con elementos distribuidos normalmente. La

matriz Sigma se calcula como el producto de x y su transpuesta, y los vectores propios (Q.S) de Sigma se obtienen para su posterior ajuste.

El algoritmo iterativo ajusta la matriz Sigma para cumplir con la condición especificada (cond.S). En un bucle, la función actualiza Sigma usando los vectores propios y los valores propios ajustados, normalizándola para que sea una matriz de correlación. Este proceso se repite hasta que la diferencia entre el ratio de los valores propios más grandes y más pequeños se aproxime a cond.S o se alcance el número máximo de iteraciones.

Finalmente, la función devuelve la matriz de correlación Sigma, proporcionando una estructura de correlación aleatoria que satisface la condición deseada.

7.5.7. Función GenAtip

La séptima función GenAtip (ver Apéndice G) está diseñada para generar observaciones multivariadas, algunas de las cuales están contaminadas, permitiendo así la simulación de datos con y sin outliers (Peña y Prieto, 2020).

La función acepta varias entradas: n.x (número de observaciones), p.x (dimensión de cada observación), par.lst (lista de parámetros del modelo), y sim.mode (modo de simulación, con un valor por defecto de 1). La salida de la función incluye x, las observaciones generadas, y lbl, etiquetas que indican si una observación es un outlier (1) o no (0).

Dependiendo del modo de simulación seleccionado, la función GenAtip sigue distintos procedimientos para generar los datos. En el Modo 1, se genera un conjunto de observaciones con una proporción específica de outliers, determinados por los parámetros alpha (proporción de outliers), delta (desviación de los outliers), y lambda (factor de escala para los outliers). El Modo 2 crea observaciones con un tipo de contaminación especificada como A, utilizando clusters limpios y contaminados con matrices de correlación y medias desplazadas. El Modo 3, similar al Modo 2, genera observaciones con otro tipo de contaminación (B), dividiendo las observaciones en más grupos contaminados con diferentes desplazamientos en las medias.

Independientemente del modo de simulación, la función sigue etapas comunes: asigna etiquetas a las observaciones (0 para no contaminadas y 1 para outliers), genera matrices de correlación para los datos limpios y contaminados, y utiliza la función mvrnorm para crear las observaciones multivariadas según las distribuciones especificadas.

En resumen, GenAtip es una herramienta flexible para la generación de datos multivariados simulados que incluye tanto observaciones normales como outliers, facilitando el estudio y la evaluación de métodos de detección de anomalías.

7.5.8. Código Main

El codigo Main (ver Apéndice H) se utiliza para ejecutar una simulación que genera datos contaminados basados en distintos modelos de contaminación y parámetros específicos (Ortiz y Londoño, 2023).

Inicialmente se definen los parámetros de la simulación, como el modo de simulación (mode.sim), que puede variar entre 1 y 3 para seleccionar diferentes modelos de outliers.

Otros parámetros incluyen alpha para el porcentaje de contaminación, delta para la distancia de los outliers, y lambda, que se utiliza solo en el método 1. Además, se especifican el número de observaciones (n.x) y la dimensión de cada observación (p.x). Estos parámetros son esenciales para controlar cómo se generan los datos contaminados.

Finalmente, se utiliza la función GenAtip para generar los datos contaminados basados en los parámetros definidos. La función produce un conjunto de datos (x.val) que contiene las observaciones simuladas. En resumen, el código realiza una simulación de datos multivariantes contaminados y prepara estos datos para un análisis posterior del sesgo.

7.5.9. Código Simulations

El codigo Simulations (ver Apéndice I) está diseñado para ejecutar simulaciones que evalúan varios métodos de detección de valores atípicos en datos multivariantes contaminados (Ortiz y Londoño, 2024).

En la primera sección, se configuran los paquetes necesarios para la simulación, asegurándose de que estén instalados y cargados en el entorno de trabajo.

En la segunda sección, se configuran los parámetros para las simulaciones, que incluyen el número de iteraciones, niveles de contaminación (alpha), dimensiones de los datos, cantidad de datos, distancias de outliers y concentración de outliers.

La tercera sección del código ejecuta las simulaciones, generando datos contaminados con la función GenAtip y aplicando métodos de detección de outliers como RSP-SAO y SAO. Se evalúan los resultados mediante matrices de confusión, calculando la precisión de los métodos de detección en identificar outliers.

En resumen, este código realiza simulaciones paralelas para evaluar la efectividad de diferentes métodos de detección de valores atípicos en datos multivariantes contaminados, almacenando los resultados para su análisis futuro.

7.6. Evaluación

Con el objeto de comparar y evaluar los resultados de la metodología propuesta se emplean tasas de detección, las cuales son métricas fundamentales en la evaluación de algoritmos y métodos de detección de outliers en conjuntos de datos

Tasa de detección real (c), también conocida como sensibilidad o tasa de verdaderos positivos, c indica la proporción de outliers reales que son correctamente identificados por el método evaluado. En términos simples, c mide la capacidad del método para capturar y detectar correctamente observaciones anómalas dentro del conjunto de datos. Una tasa de detección real alta indica que el método es efectivo en identificar la mayoría de los outliers presentes en el conjunto de datos.

Tasa de detección falsa (f) conocida como tasa de falsos positivos, f indica la proporción de observaciones normales que son incorrectamente identificadas como outliers por el método evaluado. Esta métrica es crucial porque muestra cuántas observaciones no anómalas son incorrectamente clasificadas como anómalas. Valores altos de f indican que

el método tiende a etiquetar erróneamente observaciones normales como outliers, lo cual puede ser problemático ya que introduce ruido o sesgo en el análisis de los datos.

Equilibrio y evaluación. En la evaluación de métodos de detección de outliers, es importante considerar tanto c como f para tener una comprensión completa del rendimiento del método. Idealmente, se busca maximizar c (alta sensibilidad) para capturar la mayoría de los outliers sin comprometer f (baja tasa de falsos positivos), lo que asegura que las conclusiones basadas en la identificación de outliers sean confiables y precisas.

En resumen, las tasas de detección real (c) y detección falsa (f) son indicadores clave para evaluar la precisión y eficacia de los métodos de detección de outliers. c muestra qué tan bien el método identifica los outliers verdaderos, mientras que f indica cuántas observaciones normales son incorrectamente consideradas como outliers. Un buen método de detección de outliers debería tener una alta c y una baja f para ofrecer resultados precisos y útiles en aplicaciones prácticas.

8. Resultados.

Considere una variable aleatoria p-dimensional $\boldsymbol{z} = (1-\alpha)\boldsymbol{z_c} + \alpha\boldsymbol{z_0}$ que proviene de un Modelo de Contaminación Totalmente Dependiente (Alqallaf et al., 2009), dado como una combinación de distribuciones normales multivariadas de la forma $\boldsymbol{z_c} = \mathrm{N}_p(\mathbf{0}, \mathbf{I})$ y $\boldsymbol{z_0} = \mathrm{N}_p(\delta\mathbf{1}, \lambda\mathbf{I})$, donde $\lambda \neq 0$ es un escalar de ubicación del centro multivariado de la muestra contaminada; y $\delta > 0$ denota un factor de compresión/expansión de las varianzas marginales de la contaminación. Para este experimento se establece $p = 10{,}20{,}40$, $n = 10$, $\lambda = 0.5$, $\delta = 5{,}7$ y $\alpha = 0.1{,}02{,}03{,}04$, además, se han generado $m = 50$ repeticiones aleatorias. Se compara la propuesta RSP-SAO con el SAO estándar.

En el contexto del experimento, RSP-SAO muestra valores de c superiores a 0.85 en todas las configuraciones probadas. Esto significa que más del 85% de los outliers generados según el modelo de contaminación fueron correctamente identificados por RSP-SAO. Este alto valor sugiere que el método propuesto es eficaz para identificar observaciones anómalas en comparación con el estándar SAO.

En el experimento, se observó que RSP-SAO muestra valores altos de f cuando α = 0.1 y 0.2. Valores altos de f pueden ser problemáticos porque indican que hay muchas observaciones normales que están siendo etiquetadas erróneamente como outliers. Esto puede deberse a una sobreestimación del método en la identificación de puntos extremos o anómalos, lo cual puede introducir sesgos en el análisis posterior de los datos.

A partir de los resultados presentados, se concluye que RSP-SAO tiene un desempeño robusto en la detección de outliers representados por la tasa c, superando consistentemente al método SAO estándar en todas las configuraciones evaluadas. Sin embargo, es crucial abordar los altos valores de f cuando se utiliza RSP-SAO, ya que estos pueden afectar la interpretación de los datos al identificar incorrectamente observaciones normales como outliers. Una estrategia para mitigar este problema podría ser la implementación de técnicas de ponderación o ajuste para equilibrar la precisión en la detección de outliers y la minimización de falsos positivos.

En resumen, mientras que RSP-SAO muestra fortalezas en la identificación precisa de outliers (c), la gestión adecuada de la tasa de detección falsa (f) es crucial para mejorar la confiabilidad y la aplicabilidad de la metodología propuesta en diversas aplicaciones prácticas.

Tabla No. 2. Valores c y f para todas las configuraciones de simulación.
(Los resultados se presentan en rojo para RSP-SAO y en negro para. Al final de la tabla están los valores promedio de c y f.)

ρ	α	δ	RSP-SAO		SAO	
			c	f	c	f
10	0.1	5	1,00	0.18	0.48	0,00
		7	1,00	0.18	0.72	0,00
	0.2	5	0.94	0.12	0,00	0.01
		7	0.95	0.12	0.01	0,00
	0.3	5	0.93	0.06	0,00	0.02
		7	0.92	0.07	0,00	0.01
	0.4	5	0.87	0.04	0,00	0.01
		7	0.90	0.03	0,00	0.01
20	0.1	5	1,00	0.17	0.38	0,00
		7	1,00	0.19	0.77	0,00
	0.2	5	0.93	0.12	0,00	0.01
		7	0.95	0.12	0,00	0.01
	0.3	5	0.96	0.06	0,00	0.01
		7	0.92	0.07	0,00	0.01
	0.4	5	0.91	0.03	0,00	0.01
		7	0.94	0.02	0,00	0.01
30	0.1	5	1,00	0.18	0.07	0.01
		7	1,00	0.17	0.18	0,00
	0.2	5	0.98	0.11	0,00	0.01
		7	0.97	0.12	0,00	0.01
	0.3	5	0.95	0.06	0,00	0.01
		7	0.96	0.05	0,00	0.01
	0.4	5	0.95	0.02	0,00	0.01
		7	0.94	0.02	0,00	0.01
Average			**0.95**	**0.10**	**0.11**	**0.01**

Fuente: Elaboración propia.

9. Conclusiones.

El algoritmo Stahel-Donoho (SD) es robusto y eficiente para la detección de valores atípicos multivariados debido a su capacidad para calcular medidas de periferia en múltiples direcciones proyectadas. Sin embargo, su alto esfuerzo computacional debido a la necesidad de un gran número de direcciones aleatorias limita su aplicabilidad en conjuntos de datos de alta dimensionalidad.

A diferencia de SD, el método Skewness Tighted Outlyingness (SAO) se centra en la asimetría multivariada para identificar valores atípicos, adaptándose mejor a distribuciones sesgadas. Sin embargo, su dependencia de 250p direcciones aleatorias representa un desafío significativo, ya que la utilidad de estas direcciones disminuye con el aumento de la dimensión p.

La detección de valores atípicos mediante direcciones aleatorias es inherentemente limitada, ya que no garantiza la captura efectiva de anomalías univariadas. Esto destaca la necesidad de métodos más sofisticados que utilicen direcciones de proyección más informativas y estratégicas para mejorar la precisión y eficiencia en la detección de valores atípicos.

La propuesta de utilizar direcciones de proyección más informativas mediante un método de muestreo estratificado sobre una dirección semilla ofrece una alternativa prometedora al enfoque aleatorio de SAO. Este enfoque podría potencialmente mejorar la sensibilidad y la precisión del método al adaptarse mejor a la estructura y la asimetría específica de los datos analizados.

La eficacia de los métodos de detección de valores atípicos depende en gran medida de cómo manejan la complejidad de las distribuciones de datos multivariados. Métodos como SAO muestran adaptabilidad a diferentes niveles de asimetría, lo cual es crucial para identificar anomalías en entornos no normales o sesgados.

RSP-SAO demostró una capacidad destacada con valores de c superiores a 0.85 en todas las configuraciones evaluadas. Esto indica que más del 85% de los outliers generados según el modelo de contaminación fueron correctamente identificados por RSP-SAO. Este alto rendimiento subraya la eficacia del método propuesto para la detección precisa de observaciones anómalas en comparación con SAO estándar.

Se observó que RSP-SAO exhibe valores altos de f cuando α = 0.1 y 0.2, indicando una propensión a etiquetar erróneamente observaciones normales como outliers. Este hallazgo sugiere que el método puede sobreestimar la presencia de puntos extremos, lo cual podría introducir sesgos en el análisis posterior de los datos al aumentar la tasa de detección falsa.

En conjunto, los resultados indican que RSP-SAO ofrece un desempeño robusto en la detección de outliers, superando consistentemente a SAO estándar en todas las configuraciones evaluadas según la tasa c. Sin embargo, es crucial abordar los desafíos asociados con la alta tasa de detección falsa (f) para mejorar la confiabilidad y la interpretación de los resultados de detección de outliers.

Se recomienda explorar técnicas que puedan mitigar los altos valores de f en RSP-SAO, como la implementación de estrategias de ponderación o ajuste. Estas técnicas podrían

ayudar a equilibrar la precisión en la detección de outliers y reducir la identificación incorrecta de observaciones normales como anómalas.

Es esencial validar la metodología RSP-SAO en diversas aplicaciones prácticas y configuraciones de datos para evaluar su robustez y generalización. Esta validación podría proporcionar hallazgos adicionales sobre la adaptabilidad y el rendimiento del método en diferentes contextos de análisis de datos multivariantes.

En futuras investigaciones, sería beneficioso explorar aún más la optimización de métodos de detección de valores atípicos que reduzcan la dependencia de direcciones aleatorias y mejoren la capacidad para identificar anomalías multivariadas de manera eficiente y precisa en grandes conjuntos de datos.

10. Bibliografía.

Alqallaf, F., Van Aelst, S., Yohai, V. J., y Zamar, R. H. (2009), "Propagation of Outliers in Multivariate Data", *The Annals of Statistics*, 37, 311- 331.

belenstgo.bigpress.net. (2019). *Asimetría.* Tomado de https://belenstgo.bigpress.net/texto-diario/mostrar/1609381/asimetria

Benítez, E. (2021). *Diagramas Cajas y Bigotes.* Tomado de https://matesnoaburridas.wordpress.com/2021/03/28/diagramas-cajas-y-bigotes/

Brys, G., Hubert, M., y Struyf, A. (2004), "A Robust Measure of Skewness", Journal of *Computational and Graphical Statistics*, 13, 996 - 1017.

Donoho, D. (1982), *Breakdown Properties of Multivariate Location Estimators*, Technical report, Harvard University, Boston.

Gervini, D. (2002), "The Influence Function of the Stahel-Donoho Estimator of Multivariate Location and Scatter", *Statistics & Probability Letters*, 60, 425 - 435.

Juan, J. y Prieto, F. J. (1995). "A Subsampling Method for the Computation of Multivariate Estimators with High Breakdown Point", *Journal of Computational and Graphical Statistics*, 4:4, 319 - 334.

Hubert, M. y Van der Veeken, S. (2008), "Outlier Detection for Skewed Data", *Journal of Chemometrics*, 22, 235 - 246.

Hubert, M. y Vandervieren, E. (2008), "An Adjusted Boxplot for Skewed Distributions", *Computational Statistics & Data Analysis*, 52, 5186 - 5201.

Loperfido, N. (2018), "Skewness-Based Projection Pursuit: A Computational Approach", *Computational Statistics & Data Analysis*, 120, 42 - 57.

Maronna, R. A. y Yohai, V. J. (1995), "The Behavior of the Stahel-Donoho Robust Multivariate Estimator", *Journal of the American Statistical Association*, 90, 330 – 341.

Martín, B. de U. (2022). *Los valores atípicos, anómalos o Outliers en estadística: ¿Qué son, por qué aparecen y cómo controlarlos?.* Tomado de https://trabajofinal.es/valores-atipicos-outliers-estadistica/#google_vignette

Ortiz, S. (2019), *Multivariate Outlier Detection and Robust Estimation Using Skewness and Projections*, Master's thesis, Universidad EAFIT, Medellín, Colombia.

Peña, D. y Prieto, F. J. (2001), "Multivariate Outlier Detection and Robust Covariance Matrix Estimation", *Technometrics*, 43, 286 - 300.

Peña, D. y Prieto, F. J. (2007), "Combining Random and Specific Directions for Outlier Detection and Robust Estimation in High-Dimensional Multivariate Data", *Journal of Computational & Graphical Statistics*, 16, 228 - 254.

Rousseeuw, P. J. y Croux, C. (1993), “Alternatives to the Median Absolute Deviation”, *Journal of the American Statistical Association*, 88, 1273 - 1283.

Stahel, W. A. (1981), *Robuste Schatzungen: Infinitesimale Optimalitat und Schatzungen von Kovarianzmatrizen*, Ph.D. thesis, ETH, Zurich, Switzerland.

Van Aelst, S., Vandervieren, E. & Willems, G. (2011). “Stahel-Donoho estimators with cellwise weights”, *Journal of Statistical Computation and Simulation*, 81:1, 1 - 27.

Van Aelst, S., Vandervieren, E. & Willems, G. (2012). “A Stahel–Donoho estimator based on huberized outlyingness”, *Computational Statistics and Data Analysis* 56, 531 - 542.

Van Aelst, S. (2016). “Stahel–Donoho estimation for high-dimensional data”, *International Journal of Computer Mathematics*, 93:4, 628 - 639.

Apéndice A

Función Sk.Child en R

```
################### Function Sk.Child

Sk.Child = function(X, Proy){
  #
  # 'Sk.Child' computes random and specific directions from the maximum
skewness direction
  # of a multivariate dataset, making use of the maximum projection
  # method.
  #
  # Required input arguments:
  #     X : Matrix of size 'n*p' with the multivariate data; observations
  #     by rows and variables by columns ('X' must be standardized).
  #     Proy: Maximum skewness proyection
  #
  # Outputs:
  #     chil.proy : Skewness-based random directions children
  # Authors: Santiago Ortiz (santiagoortiz00@usc.edu.co -
sortiza2@eafit.edu.co - saortizar@unal.edu.co - santy_ortiz@hotmail.com)
  # Jose Londoño (josclogo@gmail.com-jose.londono94076@u.icesi.edu.co)
  # Date: 07/2023
  p = ncol(X)
  perc = as.vector(quantile(Proy, probs = c(0.25, 0.75)))
  rand.low = sample(which(Proy <= perc[1]), 2*p)
  rand.upp = sample(which(Proy >= perc[2]), 2*p)
  sem.low = X[rand.low,]
  sem.upp = X[rand.upp,]
  chil.proy = c()#matrix(0 ,4*(p^2), p)
  for (i in 1:(2*p)) {
    ch1 = matrix(rep(sem.low[1,], 2*p), 2*p, p, byrow = T) - sem.upp
    ch.norm = matrix(rep(sqrt(apply(ch1^2, 1, sum)), 2*p),  2*p, p, byrow
= F)
    ch1.unit = ch1 / ch.norm
    chil.proy = rbind(chil.proy, ch1.unit)
  }
  return(t(chil.proy))
}
```

Apéndice B

Función adj.outly en R

```
################### Function adj.outly

adj.outly = function(X){
  #
  # 'adj.outly' computes the adjusted outlyingness measure
  # of a univariate dataset.
  #
  # Required input arguments:
  #     X : Vector of size 'n' with the univariate data.
  #
  # Outputs:
  #     Adj.AO : Outlyingness measure for each observation
  # Authors: Santiago Ortiz (santiagoortiz00@usc.edu.co -
sortiza2@eafit.edu.co - saortizar@unal.edu.co - santy_ortiz@hotmail.com)
  # Jose Londoño (josclogo@gmail.com-jose.londono94076@u.icesi.edu.co)
  # Date: 07/2023
  mu.rob = median(X)
  mc = medcouple(X)
  iqr = IQR(X)
  n = length(X)
  Adj.AO = rep(0, n)
  if (mc >= 0){
    w1 = quantile(X, 0.25) - (1.5*exp(-4*mc)*iqr)
    w2 = quantile(X, 0.75) + (1.5*exp(3*mc)*iqr)
  } else {
    w1 = quantile(X, 0.25) - (1.5*exp(-3*mc)*iqr)
    w2 = quantile(X, 0.75) + (1.5*exp(4*mc)*iqr)
  }
  for (i in 1:n) {
    if (X[i] > mu.rob){
      Adj.AO[i] = (X[i]- mu.rob) / (w2 - mu.rob)
    } else {
      Adj.AO[i] = (mu.rob - X[i]) / (mu.rob - w1)
    }
  }
  return(Adj.AO)
}
```

Apéndice C

Función Adj.Outlier en R

```
################### Function Adj.Outlier

Adj.Outlier = function(X1, type){
  #
  # 'Adj.Outlier' computes the univariate outlier detection based on
  # the Skewness Adjusted Outlyingness of a univariate dataset.
  #
  # Required input arguments:
  #     X1 : Vector of size 'n' with the univariate data.
  #     type: scalar, "1" for Medcouple calibration, otherwise
traditional upper whisker
  #
  # Outputs:
  #     lab.out : Binary, "1" Outlier, "0" Non-outlier
  # Authors: Santiago Ortiz (santiagoortiz00@usc.edu.co -
sortiza2@eafit.edu.co - saortizar@unal.edu.co - santy_ortiz@hotmail.com)
  # Jose Londoño (josclogo@gmail.com-jose.londono94076@u.icesi.edu.co)
  # Date: 07/2023
  X = X1[which(X1 <= quantile(X1, 0.6))]
  n = length(X1)
  lab.out = rep(0, n)
  mc = medcouple(X)
  iqr = IQR(X)
  p75 = quantile(X, 0.75)
  if (type == 1){
    cutoff = p75 + (1.5*exp(3*mc)*iqr)
  } else {
    cutoff = p75 + (1.5*iqr)
  }
  outliers = which(X1 >= cutoff)
  lab.out[outliers] = 1
  return(lab.out)
}
```

Apéndice D

Función maxskew en R

```
################### Function maxskew

max_skew <- function(X) {

   #
   # max_skew computes the maximum skewness projection of a multivariate
   # dataset. This function requires the function 'val_skew.m' and the
   # R library 'mrfDepth' (function 'medcouple')
   #
   # Required input arguments:
   #     X : Matrix of size 'n*p' with the multivariate data.
observations
   #     by rows and variables by columns.
   #
   # Outputs:
   #     dir_vec : Vector of size 'p*1' that maximize the skewnees
coefficient
   #     of the data in x_in.
   #
   #     skew_val : Value of the third standardized central moment of the
   #     projection of x_in in dir_vec.
  # Authors: Santiago Ortiz (santiagoortiz00@usc.edu.co -
sortiza2@eafit.edu.co - saortizar@unal.edu.co - santy_ortiz@hotmail.com)
  # Jose Londoño (josclogo@gmail.com-jose.londono94076@u.icesi.edu.co)
  # Date: 07/2023

   ## Initial verification

   if (missing(X)) stop("Input argument X is undefined")

   tol = 1.0e-10

   X.dm = dim(X)
   n.x = X.dm[1]
   p.x = X.dm[2]

   x = scale(X)                    # Standardized Data
   x1 = t(x)                       # Transposed Data

   ## Projection vector's Initializer - Taken from:
   ##    Peña & Prieto (2001), Multivariate Outlier Detection and Robust
Covariance
   ##    Matrix Estimation

   # Normalizing data by columns
   uv = apply(x*x,1,sum)
   uw = 1.0/(tol + sqrt(uv))
   uu = x * uw

   # Computing the Principal Components of normalized data in uu
```

```
   Sue = eigen(cov(uu))
   V = Sue$vectors

   # Checking which eigenvalue has the biggest absolute third centered
moment
   r = matrix(0,1,p.x)
   for (i in 1:p.x) {
      d.i = as.matrix(V[,i])
      r[i] = val_skew(x,d.i)
   }

   # Eigenvector with the biggest third moment, this is the INITIALIZER
   ik = order(abs(r))
   d0 = as.matrix(V[,ik[p.x]])
   if (val_skew(x,d0) < 0) d0 = -d0

   ## Search Algorithm
   d1 = matrix(0,p.x,1)                # Dummy vector to initialize
   cont = 0                            # Counter to stop loop if the
skewness does not converge
   # Initialize direction before the loop starts, using the vector d0 as
the best projection vector
   best_dir = d0
   tolerance = 1e-4                    # Tolerance level to stop the loop

   ## We need this for the corrected procedure
   S.sqr = pracma::sqrtm(cov(x))
   S.sqri = S.sqr$Binv

   while (norm(d0 - d1) > tolerance) {
      cont = cont + 1

      # Skewness matrix
      w.vec = t(x1) %*% d0
      aux.vec = c(w.vec) * t(x1)
      M_D = x1 %*% aux.vec
      ## Corrected procedure
      M_D = S.sqri %*% M_D %*% S.sqri

      eig.v = eigen(M_D)                        # Eigenvalues and
Eigenvectors of matrix M_D
      ind = order(abs(eig.v$values))            # Sort the eigenvalues in
ascending order
      eig.vec.sorted = eig.v$vectors[,ind]      # Ordered eigenvectors,
greatest eigenvector in the last column

      d1 = d0
      d0 = S.sqri %*% as.matrix(eig.vec.sorted[,p.x])    # New projection
vector in the loop
      d0 = d0/norm(d0,type="F")

      if (cont == 1) best_min = d0
      if ((cont >= 2) && (abs(medcouple(x %*% d0)) < abs(medcouple(x %*%
best_min)))) best_min = d0
      # Saving the vector which has the biggest third central moment
      if (abs(medcouple(x %*% d0)) > abs(medcouple(x %*% best_dir)))
best_dir = d0
```

```
      if (cont == 200) break         # If in 200 iterations there's not
convergence, the loop breaks
   }

   # Selecting the best projection vector
   dir_saved = d0

   x.bd = X %*% best_dir
   x.bm = X %*% best_min
   x.ds = X %*% dir_saved
   am.bd = abs(medcouple(x.bd))
   am.bm = abs(medcouple(x.bm))
   am.ds = abs(medcouple(x.ds))
   as.bd = abs(skewness(x.bd))
   as.bm = abs(skewness(x.bm))
   as.ds = abs(skewness(x.ds))

   if ((am.bd > am.bm && am.bd > am.ds) || (as.bd > as.bm && as.bd >
as.ds)) {
      best_of_all = best_dir
      x.ba = x.bd
   }
   else if (am.bm > am.bd && am.bm > am.ds) {
      best_of_all = best_min
      x.ba = x.bm
   }
   else {
      best_of_all = dir_saved
      x.ba = x.ds
   }

   if (skewness(x.ba) > 0) dir_vec = best_of_all
   else dir_vec = -best_of_all

   # Computing the third standardized central moment of the projection of
'X' in 'dir_vec'
   skew_v = val_skew(X,dir_vec)

   # Return values
   rval = list(dv = dir_vec, sv = skew_v)
   return(rval)
}
```

Apéndice E

Función val_skew en R

```
################### Function val_skew

val_skew <- function(x,d,km = 3) {

   #
   #     mc = val_skew(x,dir,k)
   #
   # Evaluate the moment coefficient of order k
   # for the univariate projection of multivariate data
   #
   # Inputs:  x, observations (by rows)
   #          dir, projection direction
   #          k, order of the moment (k=3 by default)
   # Output:  mc, value of the moment coefficient for the
   #          univariate data
   #
   #
   # Daniel Pena/Francisco J Prieto 23/5/00

   vs.xd = dim(x)
   vs.n = vs.xd[1]
   vs.p = vs.xd[2]

   vs.dd = dim(d)
   vs.p1 = vs.dd[1]

   if (vs.p != vs.p1) stop("Data dimensions are not correct")

   vs.t = x %*% d
   vs.tm = mean(vs.t)
   vs.tt = abs(vs.t - vs.tm)
   vs.vr = sum(vs.tt^2)/(vs.n-1)
   vs.kr = sum(vs.tt^km)/vs.n
   vs.mc = vs.kr/vs.vr^(km/2)

   return(vs.mc)
}
```

Apéndice F

Función gen_rcorr en R

```
################### Function gen_rcorr

gen_rcorr <- function(cond.S,p.x) {

  # Generate random correlation matrix.
  #   R = random_correlation_generator(p, iteration_times) is a function
  #   that generate a random correlation estructure for a p-dimensional
  #   multivariate random variable. The random correlation structure
  #   is described in [1].
  #
  #[1] Agostinelli, C., Leung, A., Yohai, V.J., Zamar, R.H., 2015. Robust
  #   estimation of multivariate location and scatter in the presence of
  #   cellwise and casewise contamination. TEST 24(3), 441-461.
  #   INPUTS:
  #   p: Is the dimension of the normal multivariate random variable.
  #   iteration_times (Not required): Is the maximum iteration times for
  #   the convergence of the algorithm to estimate R, default 99 times.
  #   OUTPUTS:
  #   R: Is the correlation matrix that contain the random correlation
  #   estructure generated.
  #   Autor: Henry G. Velasco (hgvelascov@eafit.edu.co)

  maxits = 100
  tol = 1.0e-5

  if (p.x < 3) stop("p must be larger than 2")

  lambda = sort(c(1,runif(p.x-2,1,cond.S),cond.S))
  x = matrix(rnorm(p.x*p.x),p.x,p.x)
  Sigma = x %*% t(x)

  S.eig = eigen(Sigma)
  Q.S = S.eig$vectors
  ratio = 0
  iter = 0

  while ((abs(ratio - cond.S) > tol) && (iter < maxits)) {
    iter = iter + 1
    Sigma = Q.S %*% diag(lambda) %*% t(Q.S)
    Sig.sr = 1.0/sqrt(c(diag(Sigma)))
    Sigma = t(Sig.sr * Sigma) * Sig.sr
    S.eig = eigen(Sigma)
    Q.S = S.eig$vectors
    lambda = S.eig$values
    ratio = lambda[1]/lambda[p.x]
    lambda[p.x] = lambda[1]/cond.S
  }

  return(Sigma)
}
```

Apéndice G

Función GenAtip en R

```
################### Function GenAtip

GenAtip <- function(n.x,p.x,par.lst,sim.mode = 1) {

  #
  # Generate contaminated observations controlling the parameters
  #
  #     GenAtip(n.x,p.x,par.lst,sim.mode)
  #
  # Inputs:  n.x,  number of observations
  #          p.x,  dimension of each observation
  #          par.lst, parameters to use in the data generation model
  #          sim.mode, = 1 , normal observations and contamination
  # Output:  x, observations
  #          lbl, labels of the observations (lbl = 0 for an
uncontaminated observation
  #                                          lbl = 1 for an outlier)
  #

  #  Daniel Pena / Francisco J Prieto 06/03/2020

  ## Labels for outliers and central observations

  lbl = matrix(1,n.x,1)

  ## Generate outliers according to the selected pattern

  if (sim.mode == 1) {

    ## Usual contamination model (1 group of outliers)

    ## Parameters

    alpha = par.lst[1]
    n.1 = floor(n.x*alpha)
    n.0 = n.x - n.1

    lbl[1:n.0] = 0

    ## Centers of clusters

    delta = par.lst[2]            # Default value = 1
    m.0 = matrix(0,1,p.x)
    m.1 = delta*matrix(2,1,p.x)

    ## Data for each cluster

    lambda = par.lst[3]
    x = matrix(rnorm(n.x*p.x),n.x,p.x)
    x[1:n.0,] = scale(x[1:n.0,])
```

```
    x[(n.0+1):n.x,] = scale(x[(n.0+1):n.x,])
    x[(n.0+1):n.x,] = sqrt(lambda)*x[(n.0+1):n.x,] + matrix(1,n.1,1) %*%
m.1

  } else if (sim.mode == 2) {

    ## Contamination type A from HL and SO

    cn.mx = 100
    cn.mn = 1
    dst.mx = 10
    dst.mn = 5

    n.gr = 2

    alpha = par.lst[1]
    n.cnt.g = round(n.x*alpha/n.gr)
    n.cln = n.x - n.gr*n.cnt.g

    delta = par.lst[2]        # Default value = 1
    lambda = par.lst[3]       # Default value = 1

    lbl[1:n.cln] = 0

    ## Observations in each cluster

    Cov.cln = gen_rcorr(cn.mx,p.x)
    mn.cln = matrix(0,p.x,1)
    x.cln = mvrnorm(n.cln,mn.cln,Cov.cln)

    Cov.cnt = lambda*diag(p.x)
    v.1 = runif(p.x-1,dst.mn,dst.mx)
    v.2 = -runif(p.x-1,dst.mn,dst.mx)
    mn.cnt.1 = delta*matrix(c(v.1,0),p.x,1)
    mn.cnt.2 = delta*matrix(c(v.2,0),p.x,1)
    x.cnt.1 = mvrnorm(n.cnt.g,mn.cnt.1,Cov.cnt)
    x.cnt.2 = mvrnorm(n.cnt.g,mn.cnt.2,Cov.cnt)

    ## Simulated data

    x = rbind(x.cln,x.cnt.1,x.cnt.2)

  } else if (sim.mode == 3) {

    ## Contamination type B from HL and SO

    cn.mx = 100
    cn.mn = 1
    dst.mx = 10
    dst.mn = 5

    n.gr = 4

    alpha = par.lst[1]
    n.cnt.g = round(n.x*alpha/n.gr)
    n.cln = n.x - n.gr*n.cnt.g
```

```
    delta = par.lst[2]          # Default value = 1
    lambda = par.lst[3]         # Default value = 1

    lbl[1:n.cln] = 0

    ## Observations in each cluster

    Cov.cln = gen_rcorr(cn.mx,p.x)
    mn.cln = matrix(0,p.x,1)
    x.cln = mvrnorm(n.cln,mn.cln,Cov.cln)

    Cov.cnt = lambda*diag(p.x)
    v.1 = runif(p.x-1,dst.mn,dst.mx)
    v.2 = -runif(p.x-1,dst.mn,dst.mx)
    v.3 = c(matrix(c(1,-1),p.x-1,1))*runif(p.x-1,dst.mn,dst.mx)
    v.4 = c(matrix(c(-1,1),p.x-1,1))*runif(p.x-1,dst.mn,dst.mx)
    mn.cnt.1 = delta*matrix(c(v.1,0),p.x,1)
    mn.cnt.2 = delta*matrix(c(v.2,0),p.x,1)
    mn.cnt.3 = delta*matrix(c(v.3,0),p.x,1)
    mn.cnt.4 = delta*matrix(c(v.4,0),p.x,1)
    x.cnt.1 = mvrnorm(n.cnt.g,mn.cnt.1,Cov.cnt)
    x.cnt.2 = mvrnorm(n.cnt.g,mn.cnt.2,Cov.cnt)
    x.cnt.3 = mvrnorm(n.cnt.g,mn.cnt.3,Cov.cnt)
    x.cnt.4 = mvrnorm(n.cnt.g,mn.cnt.4,Cov.cnt)

    ## Simulated data

    x = rbind(x.cln,x.cnt.1,x.cnt.2,x.cnt.3,x.cnt.4)
  }

  ## Return values
  cv = list(x = x, lbl = lbl)
  return(cv)
}
```

Apéndice H

Código Main en R

```
#########################################################################
#################### Main code
#########################################################################

  # Authors: Santiago Ortiz (santiagoortiz00@usc.edu.co -
sortiza2@eafit.edu.co - saortizar@unal.edu.co - santy_ortiz@hotmail.com)
  # Jose Londoño (josclogo@gmail.com-jose.londono94076@u.icesi.edu.co)
  # Date: 12/2023

library(mrfDepth)              # Function medcouple
library(e1071)                 # Function skewness
library(pracma)                # Function findpeaks
library(MASS)                  # Function mvrnorm

#########################
#  Simulation runs
#  Choice of contamination model
#########################

# Simulation parameters

mode.sim = 2       # Simulation mode
# = 1 one oultier group
# = 2 two outlier groups (first model from Henry and Santiago)
# = 3 four outlier groups (second model from Henry and Santiago)

alpha = 0.1
delta = 1             # Default value for methods 2 and 3
lambda = 0.1          # Only used for method 1
par.lst = c(alpha,delta,lambda)

n.x = 500
p.x = 3

x.val = GenAtip(n.x,p.x,par.lst,mode.sim)
#aa = out_skew(x.val$x
```

Apéndice I

Código Simulations en R

```
#########################################################################
#################### Simulations
#########################################################################

  # Authors: Santiago Ortiz (santiagoortiz00@usc.edu.co -
sortiza2@eafit.edu.co - saortizar@unal.edu.co - santy_ortiz@hotmail.com)
  # Jose Londoño (josclogo@gmail.com-jose.londono94076@u.icesi.edu.co)
  # Date: 02/2024

set.seed(50)
## Required Packages
my_packages = c("expm","gridExtra","tidyverse","knitr","kableExtra",
                "IRdisplay","mrfDepth","e1071","pracma","MASS","mixtools",
                 "doParallel","foreach","extraDistr","rrcov","Rfast",
                 "dobin","FNN","parallel","depthTools","OutliersO3",
                   "mvoutlier","fds", "caret", "tictoc")
not_installed = my_packages[!(my_packages %in% installed.packages()[ ,
"Package"])]
if (length(not_installed)) install.packages(not_installed, dependencies =
TRUE)
for (q in 1:length(my_packages)) {
  library(my_packages[q], character.only = TRUE)}

## Working Directory
setwd(dirname(rstudioapi::getSourceEditorContext()$path))
## Load outlier detection codes and auxiliary routines
source("OL_skew_self_V2.R")

ini<-Sys.time()

iter = 100
alpha = c(0.1,0.2,0.3,0.4)
dimension = c(10,20,40)
cant_datos = 10*dimension
outl_dist = c(3,4)
outl_concent = c(0.5)

#########################################################################
####### Multivariate normal distribution N(0,I) contaminated with
(alpha/2)*N(0 + delta,lambda*I) #######
################################
registerDoParallel(detectCores())
#res4 = foreach (i2 = 1:length(dimension), .combine = rbind) %do% {
  #res3 = foreach (i1 = 1:length(alpha), .combine = rbind) %do% {
    #res2 = foreach (i3 = 1:length(outl_dist), .combine = rbind) %do% {
      #res1 = foreach (i4 = 1:length(outl_concent), .combine = rbind)%do%
             #{res0 = foreach (i5 = 1:iter, .combine = rbind) %dopar% {
resul = matrix(0, iter, 4)
resul.t = matrix(0, 0, 4)
for (i2 in 1:length(dimension)) {
```

```
  for (i1 in 1:length(alpha)) {
    for (i3 in 1:length(outl_dist)) {
      for (i4 in 1:length(outl_concent)) {
        for (i5 in 1:iter) {
          datos_limp = round(cant_datos[i2]*(1-alpha[i1]))
          datos_cont = cant_datos[i2]-datos_limp
          data.sim =
GenAtip(cant_datos[i2],dimension[i2],c(alpha[i1],outl_dist[i3],outl_conce
nt[i4]),1)
          label.X = as.factor(data.sim$lbl)
          X = data.sim$x
          ##   # pairs(X[,1:5], pch = 20)      ##

          # RSP-SAO
          Sk = max_skew(X)
          Sk.proy = X %*% Sk$dv
          # hist(Sk.proy)
          Sk.dirs = Sk.Child(X, Sk.proy)
          Skproy.Child = X %*% Sk.dirs
          #Adj.SDO = apply(apply(cbind(Skproy.Child, Sk.proy), 2,
adj.outly), 1, max)
          proy.t = cbind(Skproy.Child, Sk.proy)
          Adj.SDO = apply((proy.t - apply(proy.t, 2, median)) /
apply(proy.t, 2, Qn), 1, max)
          # hist(Adj.SDO)
          out.points = Adj.Outlier(Adj.SDO, 1)
          W0.cm = confusionMatrix(as.factor(out.points), label.X)$table
          #W0.cm

          # SAO
          W1 = as.numeric(adjOutlyingness(X)$nonOut)
          W1[which(W1 == 1)] = 2
          W1[which(W1 == 0)] = 1
          W1[which(W1 == 2)] = 0
          W1.cm = confusionMatrix(as.factor(W1), label.X)$table

          # RESULTS
          resul[i5,] = c(W0.cm[2,2]/sum(W0.cm[,2]),
W0.cm[2,1]/sum(W0.cm[,1]),
                         W1.cm[2,2]/sum(W1.cm[,2]), W1.cm[2,1]/sum(W1.cm[,1]))
        }
        resul1 = apply(resul, 2, mean)
      }
      resul.t = rbind(resul.t, resul1)
      cat(i2, i1, i3, i4)
    }
  }
}
#print("Experiment Symmetric TypeA has finished")
save(resul.t, file = "Symm.RData")
write.csv(resul.t, file = "Symm.csv")

stopImplicitCluster()
fin<-Sys.time();fin-ini
###########
### FIN ###
###########
```

Anexo A

Certificación de Ponencia Simposio

UNIVERSIDAD NACIONAL DE COLOMBIA

DEPARTAMENTO DE ESTADÍSTICA
FACULTAD DE CIENCIAS
SEDE BOGOTÁ

CERTIFICA QUE

SANTIAGO ORTIZ ARIAS

IDENTIFICADO (A) CON CC 1037640332

PARTICIPÓ EN CALIDAD DE PONENTE EN EL

32° SIMPOSIO INTERNACIONAL DE ESTADÍSTICA

CON UNA COMUNICACIÓN ORAL TITULADA "RANDOM SPECIFIC PROJECTION DIRECTIONS FOR SKEWNESS ADJUSTED OUTLYINGNESS".

APROBADO SEGÚN RESOLUCIÓN 0468 de 2023 DEL CONSEJO DE FACULTAD, POR EL CUAL SE APRUEBA Y DISTRIBUYE EL PRESUPUESTO PARA LA REALIZACIÓN DEL PROYECTO: SIMPOSIO INTERNACIONAL DE ESTADÍSTICA, Y SE DICTAN OTRAS DISPOSICIONES.

REALIZADO DEL 31 DE JULIO DE 2023 AL 4 DE AGOSTO DE 2023 EN LA CIUDAD DE IBAGUÉ, CON UNA INTENSIDAD DE 35 HORAS.

DADO EN IBAGUÉ, EL 4 DE AGOSTO DE 2023.

RAMÓN GIRALDO H.
Director
DEPARTAMENTO DE ESTADÍSTICA

HELBER DE JESÚS BARBOSA B.
Secretario
FACULTAD DE CIENCIAS