Métodos Robustos de Estimación de Correlación

Br. María Alejandra Rodríguez Jáuregui
Profesor Guía: Prof. José Luis Paredes

Proyecto de Grado presentado ante la ilustre Universidad de Los Andes
como requisito final para optar al Título de Ingeniero de Sistemas

Mérida, Venezuela
Julio 2005

UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

Índice general

Índice de Tablas

Índice de Figuras

Agradecimientos

- Al profesor José Luís Paredes por mostrarme que la inteligencia, disciplina y sencillez pueden estar en una misma persona, gracias por permitirme trabajar con usted.
- Al profesor Giorgio Bianchi del Departamento de Matemática y al profesor Juan Marcos Ramírez de la Escuela de Eléctrica, por el apoyo incondicional que me brindaron, y por compartir sus conocimientos conmigo.
- A mi familia, por ser pilar fundamental en mi educación y aprendizaje diario.
- Al CDCHT por respaldar económicamente este proyecto de investigación bajo el código I-797-04-02-F.
- A todos los que de alguna manera pusieron su granito de arena en estas líneas.

Resumen

Los métodos de procesamiento de señales hasta hoy han sido dominados por la idea de que el ruido inmerso dentro de una señal tiene un comportamiento de tipo gaussiano. Sin embargo, los ambientes industriales están caracterizados por ruidos externos de naturaleza impulsiva. Bajo estas condiciones, los métodos de estimación, basados en procesos gaussianos, fallan, surgiendo la necesidad de considerar técnicas robustas de procesamiento de señales que sean tan eficientes como sea posible en la presencia de ruido impulsivo. Los métodos basados en momentos de segundo orden son frecuentemente utilizados en la práctica para estimar la correlación que existe entre dos variables X e Y. En particular, bajo el principio de máxima verosimilitud y si ambas variables obedecen una distribución gaussiana bivariante, la correlación muestral se reduce a la sumatoria del producto de las muestras (X_i, Y_i). Sin embargo, dicha estimación no es robusta cuando las variables se alejan del modelo Gaussiano siendo caracterizadas más eficientemente por distribuciones de colas pesadas que modelan un comportamiento tipo impulsivo. Por esta razón surge la necesidad de buscar métodos más robustos ante ruidos de naturaleza impulsiva que muestren mejores desempeños que los métodos clásicos basados en la combinación lineal de los datos. En el presente trabajo se hace una comparación de los métodos robustos de estimación de correlación, tomando como base el método de correlación tradicional y evaluando nuevos métodos relacionados con estimadores robustos ante ruido impulsivo. En particular se compara el desempeño del método propuesto por Arce y Li (2002) [1] con respecto al método tradicional de cálculo de correlación y un método correlación propuesto por Rousseaw (1999) [2]. A fin de ilustrar el desempeño de estas teorías de correlación, se aplican estos conceptos a distintos algoritmos que dependen notablemente de estimaciones de correlación muestral, tal como el criterio de la Descripción de la Minima Longitud (MDL), Clasificación

de Señales Múltiples (MUSIC), y la Normalización de Datos de Microarray de ADN Complementario.

Palabras claves: *Estimación de Correlación, Matrices Robustas de Correlación, Robusto, Ruido Impulsivo, Dependencia de Correlación.*

Lista de Términos

- $\bar{\beta}$: Parámetro de localización estimado bajo el modelo Gaussiano
- $\tilde{\beta}$: Parámetro de localización estimado bajo el modelo Laplaciano
- $Cov(X, Y)$: Covarianza entre X e Y
- $F(X_1, X_2, ..., X_n)$: Función de Probabilidad Conjunta
- M_{ij} : Momento de Orden $i + j$
- $\bar{R}_{XY}$: Correlación Tradicional entre X e Y
- $\bar{R}_{XX}$: Autocorrelación Tradicional de X
- $\tilde{R}_{XY}$: Correlación Mediana Muestral entre X e Y
- $\tilde{R}_{XX}$: Autocorrelación Mediana Muestral de X
- $\hat{R}_{XY}$: Covarianza Mediana Muestral entre X e Y
- $\hat{R}_{XX}$: Autocovarianza Mediana Muestral de X
- S: Covarianza con Determinante Minimo
- $d(X_i)$: Distancia Robusta

Capítulo 1

Introducción

El procesamiento y análisis de señales es una parte integral de numerosas áreas, tales como, instrumentación, automatización a nivel industrial, procesos de adquisición de datos, comunicaciones, control automático de procesos. La mayoría de los métodos de procesamiento de señales han sido dominados por la idea de que los procesos de interferencia y ruido inmerso dentro de una señal tienen un comportamiento de naturaleza gaussiana. Esto es justificado en parte, teóricamente por el teorema de límite central, y es razonable para los casos donde el ruido es interno al sistema, tales como el zumbido y estática de un radio, las oscilaciones de un sistema realimentado, fluctuaciones espontáneas de corriente o voltaje en los circuitos y elementos eléctricos o simplemente el ruido térmico introducido por el calentamiento de los componentes electrónicos. Sin embargo, los ambientes industriales están caracterizados por ruidos externos impulsivos, debido a la actividad constante de sistemas de ignición, motores eléctricos, descargas por efecto de líneas de alta tensión, sistemas de diatermia y ruido de conmutación, los cuales son sin duda procesos de naturaleza no gaussiana [3]. Bajo estas condiciones, se conoce que los métodos de procesamiento de señales elaborados bajo la suposición de contaminación de fondo es de tipo gaussiano no son eficientes, por lo tanto, es necesario considerar técnicas robustas de procesamiento de señales que puedan desempeñarse tan eficientemente como sea posible en la presencia de interferencia y ruido de naturaleza impulsiva.

1.1. Motivación

Actualmente se reconoce que existe carencia de una teoría general de robustez en importantes áreas de procesamiento de señales tal como modelado de sistemas y teoría espectral. Las áreas establecidas en el procesamiento robusto de señales, están basadas en las aplicaciones de estimación de parámetro, donde la teoría de estadística robusta satisface principalmente al muestreo que contiene data contaminada con puntos dispersos.

El presente Proyecto de Grado esta motivado por una investigación reciente desarrollada por Arce y Li [1], donde se mostró que la estimación de la correlación establecida mediante el principio de verosimilitud máxima siguiendo el modelo Laplaciano tiene una estructura basada en el operador de mediana con ponderación variable. Así como la estimación de correlación tradicional, se deduce partiendo del modelo Gaussiano, la estimación de correlación basada en el operador de mediana tiene sus orígenes en el modelo Laplaciano y presenta una estructura sorprendentemente simple. En contraste a la correlación lineal, la correlación de mediana es robusta en presencia de ruido de naturaleza impulsiva. Particularmente, las ponderaciones en este contexto no asumen valores fijos como es el caso de los filtros ponderados de mediana, sino que adquieren los valores aleatorios determinados por los datos adyacentes de los mismos. Originando así, nuevas propiedades del operador de mediana, que resultan ser útiles dentro del campo de procesamiento de señales y comunicaciones.

Considerando que los métodos introducidos por Arce y Li [1], pueden tener impacto en áreas en las cuales son necesarios algoritmos eficientes para el procesamiento de señales en ambientes impulsivos, nos propusimos hacer un estudio comparativo del desempeño de diversos métodos propuestos en [1] y el método tradicional de cálculo de correlación así como también un tercer método reportado en la literatura [4]. A fin de ilustrar el funcionamiento de estas teorías de correlación, aplicamos estos conceptos a distintos algoritmos que dependen notablemente de estimaciones de correlación muestral tal como la Descripción de la Longitud Minima (MDL) [5], Clasificación de Señales Múltiples (MUSIC) [6], Normalización de Datos de Microarray de ADN complementario [7].

1.2. Objetivos y Organización

En el presente Proyecto de Grado se propone hacer una comparación de los métodos robustos de estimación de correlación, tomando como base el método de correlación tradicional y evaluando los nuevos métodos relacionados con estimadores robustos ante ruido de naturaleza impulsiva. En detalle se desea:

- Comparar el desempeño de los métodos robustos de correlación, en particular, el método propuesto por Arce y Li con respecto al método tradicional de cálculo de correlación y otros métodos existentes en la literatura.
- Proponer alternativas que mejoren el método propuesto por Arce y Li.

La organización de este proyecto consta de 8 capítulos, los cuales se describen a continuación: El Capítulo 2 presenta el marco de referencia donde se explican los conceptos básicos para abordar el tema. El Capítulo 3 describe los conceptos del método de correlación lineal muestral. En el Capítulo 4 se presentan los fundamentos de los métodos robustos de correlación propuesto por Arce en [1], introduciendo la analogía con respecto a la correlación tradicional, los conceptos de correlación mediana y covarianza mediana muestral. En el Capítulo 5 se describe el concepto de matriz de covarianza con determinante mínimo propuesto por Rousseaw en [4]. En el Capitulo 6 se describen las aplicaciones propuestas, específicamente Descripción de la Longitud Minima (MDL) [8], Clasificación de Múltiples Señales (MUSIC) [6], Normalización de Datos de Array de ADN complementario [7][9], y se da una breve explicación acerca de la dependencia de la matriz de correlación que tienen cada uno de estos algoritmos. En el Capítulo 7 se presentan los resultados de la simulación, comparando el desempeño de los distintos métodos en las aplicaciones propuestas. Para finalizar en el Capítulo 8 se señalan los aportes del trabajo realizado, conclusiones y recomendaciones de trabajos a futuro.

Capítulo 2

Marco De Referencia

La correlación permite unificar el estudio frecuencial de las señales deterministicas y aleatorias [10], generando junto con la transformada de Fourier el modelado y procesamiento estadístico de señales.

2.1. Correlación

La correlación es un parámetro estadístico que mide la dependencia lineal entre dos conjuntos de variables X e Y, en donde $X = \{X_1, X_2, \ldots, X_n\}$ e $Y = \{Y_1, Y_2, \ldots, Y_n\}$ [11][2][12]. La dependencia lineal entre dos conjuntos de variables se puede observar representando cada par de valores $(X_i, Y_i|_{i=1}^{n})$ de los conjuntos de datos en el plano cartesiano: cada par de valores (X_i, Y_i) da lugar a un punto en el plano y el conjunto de puntos que se resultante recibe el nombre de diagrama de dispersión o nube de puntos.

Según la forma de la nube de puntos se puede observar el grado de dependencia lineal entre las variables. Si el diagrama de puntos entre dos conjunto de datos genera una recta, existe una alta relación entre dichas variables, cuando la correlación es nula los puntos se disponen aleatoriamente en un circulo; si las varianzas de las variables son iguales ($\sigma_X = \sigma_Y$); y en forma de elipse cuando las varianzas de las variables son distintas ($\sigma_X \neq \sigma_Y$). La figura 2.1**(a)** representa el gráfico de nube de puntos de dos variables que no están correlacionadas. La figura 2.1**(b)** muestra el gráfico de nube de puntos de dos variables correlacionadas. Como la recta generada por estos puntos es

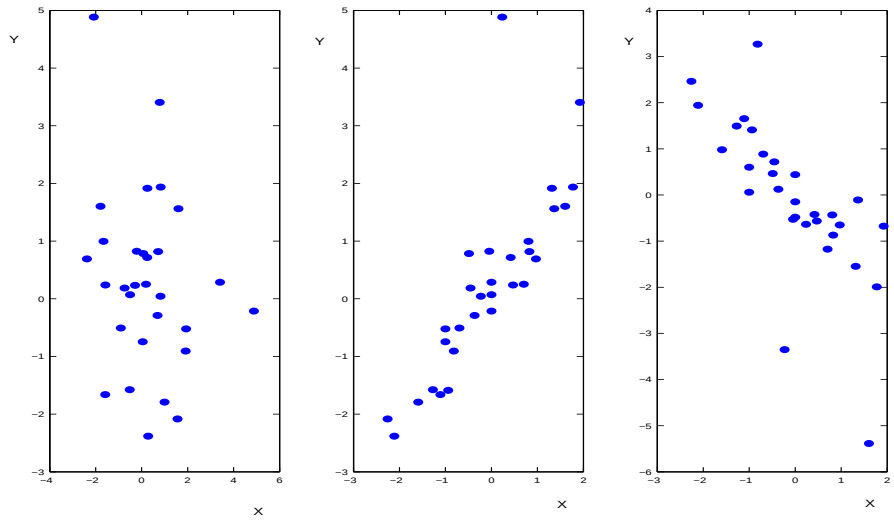

Figura 2.1: Gráfico de nube de puntos que representa la correlación entre dos variables X e Y

creciente la correlación es positiva o directa, es decir, al aumentar una variable, la otra tiene también tendencia a aumentar. La figura 2.1**(c)** representa el gráfico de nube de puntos de dos variables correlacionadas, al ser la recta decreciente la correlación es negativa o inversa, es decir, al aumentar una variable, la otra tiene tendencia a disminuir.

Adicionalmente el gráfico de nube de puntos, permite observar la posible existencia de valores alejados de la tendencia natural de los datos, sin embargo, la apreciación visual de la existencia de estos valores no proporciona medición cuantitativa del grado de dependencia entre dos variables. Se puede determinar si la correlación de los datos es fuerte o débil, positiva o negativa a través de la estimación del parámetro de correlación (ρ).

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

donde $cov(X,Y)$ es la covarianza entre las variables X e Y.

El coeficiente de correlación es un valor que fluctúa entre $-1 < \rho < 1$, donde -1 representa correlación perfecta de sentido negativo, la cual es representada en el diagrama de dispersión como una recta de pendiente negativa. Cuando la nube de puntos genera una recta de pendiente positiva esta asociado con correlación perfecta

en sentido positivo ($\rho = +1$) . Cuanto más cercano a cero sea el valor de ρ, indica una mayor debilidad de la relación entre las variables o incluso ausencia de correlación cuando ρ es igual a cero.

2.2. Justificación de Métodos Robustos de Estimación de Correlación

La mayoría de las señales en la práctica son generalmente perturbadas por señales aleatorias indeseables, que distorsionan la señal original. Estas perturbaciones se manifiestan como irregularidades dentro de la señal, y es lo que se conoce generalmente como "ruido".

Cuando el ruido que contamina la señal es de tipo gaussiano, la correlación mide el grado de similaridad entre dos señales, pero en presencia de impulsos, esta no es muy eficiente.

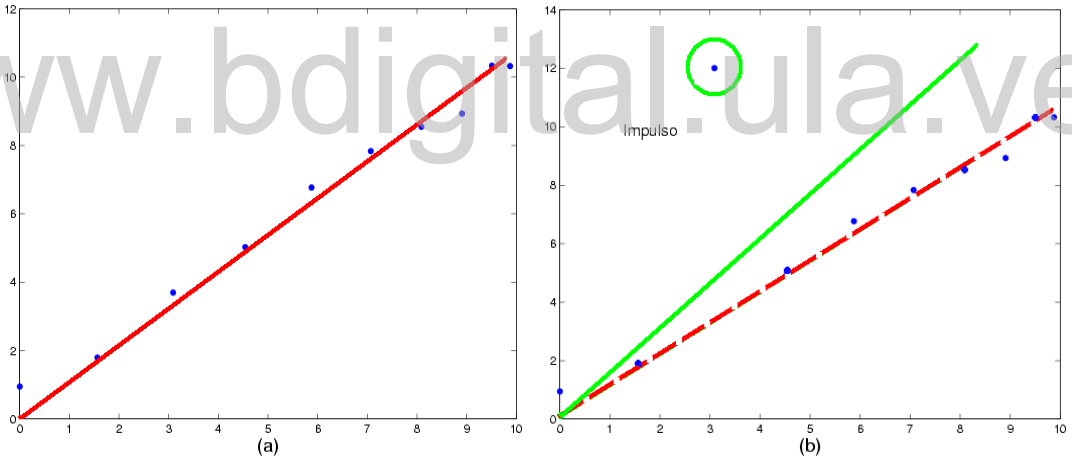

Figura 2.2: Gráfico de nube de puntos que representa la correlación entre dos variables X e Y.

Por ejemplo, considere dos variables fuertemente correlacionadas, en la figura 2.2(a), se puede observar la tendencia de los puntos a una recta en el gráfico de dispersión donde ρ se aproxima 1, para este caso especifico $\rho = 0,996$. En la figura 2.1(b) se representa las misma variables pero cambiando uno de los datos por un impulso, arrojando un valor de $\rho = 0{,}723$, valor que indica una correlación débil. Se puede decir entonces

que dos variables pudieran estar muy correlacionadas y por lo tanto acercarse a una recta en el gráfico de dispersión pero debido a la presencia de impulsos hacen que la correlación determinada como la suma ponderada de las muestras, no represente la verdadera tendencia que sigue ambas variables. De allí la importancia de considerar métodos robustos de estimación de correlación que en presencia de impulsos permitan cuantificar de manera efectiva el grado de semejanza entre distintas señales.

2.3. Modelo Gaussiano

La distribución gaussiana representa sin duda un modelo determinante en estadística e ingeniería. A pesar de la simplicidad matemática del modelo, el teorema del límite central le ha dado a las distribuciones gaussianas un lugar privilegiado en la historia de la estadística. Adicionalmente, el teorema del límite central ha sido un paradigma, favoreciendo el uso de métodos lineales incluso en las condiciones en las cuales la naturaleza no gaussiana de los procesos es evidente [3].

Conceptualmente, el teorema del límite central describe los procesos gaussianos como la superposición de muchos efectos pequeños e independientes. Éste es el caso por ejemplo del ruido térmico, que se genera como la superposición de un gran número de interacciones aleatorias independientes a nivel molecular.

Íntimamente ligado al modelo Gaussiano están los métodos de estimación lineal. Por ejemplo, dado un sistema de muestras de tipo Gaussiano independientes e idénticamente distribuidas (i.d.d), es bien sabido que la estimación óptima de localización es la media muestral.

Numerosas áreas tales como las comunicaciones, instrumentación, control automático de procesos, no se han escapado del uso del modelo Gaussiano. Aunque se han encontrado muchos procesos importantes que son definitivamente de tipo no Gaussiano, existe una gran cantidad de sistemas prácticos que todavía viven dentro del mundo Gaussiano y por lo tanto justifican la aplicación de técnicas lineales. Una preocupación seria es que, en general, un sistema diseñado asumiendo que es un proceso que sigue una distribución de tipo Gaussiano, demostrará degradaciones drásticas del funcionamiento cuando la estadística del ruido provenga de modelos de colas pesadas.

Por ejemplo, se sabe que la media muestral, siendo un estimador lineal, presenta un comportamiento óptimo para procesos Gaussianos, mientras que se convierte en un estimador perceptiblemente pobre en presencia de contaminación de tipo impulsiva.

2.4. Modelo No Gaussiano

La naturaleza impulsiva del ruido que afecta las señales encontradas en procesos industriales procede de una gran variedad de fuentes externas impulsivas (maquinaria con rotor, la ignición del motor y otros fenómenos de descarga), así como de ciertos casos de interferencia electromagnética [3]. Para modelar estos procesos, se ha propuesto una gran variedad de distribuciones de colas más pesadas que las generadas por el modelo Gaussiano, como alternativas viables a la distribución gaussiana. La mayoría de estos modelos se basan en distribuciones simétricas y localizadas en cero. La utilidad de estos modelos es determinada generalmente por la compensación entre la fidelidad y la complejidad. La fidelidad orientada al desarrollo de algoritmos más exactos y más eficientes del procesamiento de señales, mientras que la complejidad está orientada hacia modelos más simples de los cuales se pueden derivar algoritmos más manejables [3].

En estadística robusta se asume a menudo que las distribuciones "impulsivas", obedecen un modelo del tipo ε-contaminación, que esta caracterizado por el índice de error en la respuesta, y esta representado por la función de densidad:

$$f(x) = (1 - \varepsilon) f_0(x) + \varepsilon h(x)$$

donde $f_0(x)$, es la densidad nominal, y usualmente corresponde a una distribución gaussiana, ε es una constante positiva pequeña que representa el índice de contaminación, y $h(x)$ es una función de densidad arbitraria de colas pesadas la cual representa la presencia de puntos dispersos en los datos. Sin embargo, en situaciones prácticas no se conoce con exactitud el tipo de distribución que modela el proceso impulsivo($h(x)$), por esta razon es necesario considerar distribuciones más simples que representen la impulsividad del ruido.

En el procesamiento de señales, muchos fenómenos de ruido pueden modelarse como

la superposición de pequeños efectos, independientes e impulsivos. Generalizando el teorema del límite central, los procesos de esta naturaleza son modelados por una clase de distribuciones de colas pesadas con varianza infinita conocida como α-estable.

Así como el teorema del límite central ha influido para el uso del modelo Gaussiano en muchos casos prácticos es necesario el uso de los modelos α-estable, en los problemas que implican impulsividad.

El modelo α-estable ha sido empleado en una variedad de campos incluyendo hidrología, economía, física e ingeniería [13]. Recientemente, han sido el tema de mayor atención en comunicaciones y el procesamiento de señales.

2.5. Procesos α-estable

En años recientes, el interés en modelar señales ha llevado a los investigadores a la conclusión de que muchos fenómenos naturales se pueden representar mejor por distribuciones de naturaleza más impulsiva. Una excelente introducción a un tipo particular de señales no gaussianas es específicamente la distribución de tipo α-estable, la cual representa una alternativa interesante para modelar fenómenos de naturaleza impulsiva. Esto es debido a que el modelo Gaussiano no logra describir de manera satisfactoria señales que tienen un comportamiento impulsivo. Los procesos α-estables y Gaussianos son frecuentemente utilizados para modelar señales de ruido, pero se ha demostrado que la distribución α-estable tiene colas más pesadas que la distribución gaussiana, por lo tanto proporciona una aproximación mucho mejor a ciertas señales encontradas en la práctica.

La distribución α-estable, la cual puede modelar fenómenos de naturaleza impulsiva, es una generalización de la distribución gaussiana y es atractiva debido a dos propiedades básicas [13]:

- Satisface la propiedad de estabilidad, la cual indica que si X, X_1 y X_2 son variables aleatorias independientes de tipo α-estable que obedecen una misma distribución, entonces existe constantes μ_1, μ_2, v_1, v_2 tales que:

$$v_1 X_1 + v_2 X_2 \sim \mu_1 X + \mu_2$$

donde v_1, v_2, μ_1 y μ_2 son constantes y donde $\sim$ implica igualdad en la distribución.

- Satisface el teorema de límite central. Es decir, si X es α-estable, si y solo si X es el limite cuando $n \to \infty$ de la distribución de la suma:

$$S_n = \frac{X_1 + X_2 + \ldots X_n}{a_n} - b_n$$

donde $X_1, X_2, \ldots, X_n$ son las muestras i.i.d, el parámetro b_n es real y a_n es real y positivo

Una función de distribución es α-estable si su función característica tiene la forma:

$$\phi(t) = exp(-\gamma|t|^{\alpha})$$

donde los parámetros α y γ satisfacen $0 < \alpha \leq 2$ y $\gamma > 0$. El parámetro α es conocido como exponente característico y controla la pesadez de las colas en la función de la densidad.

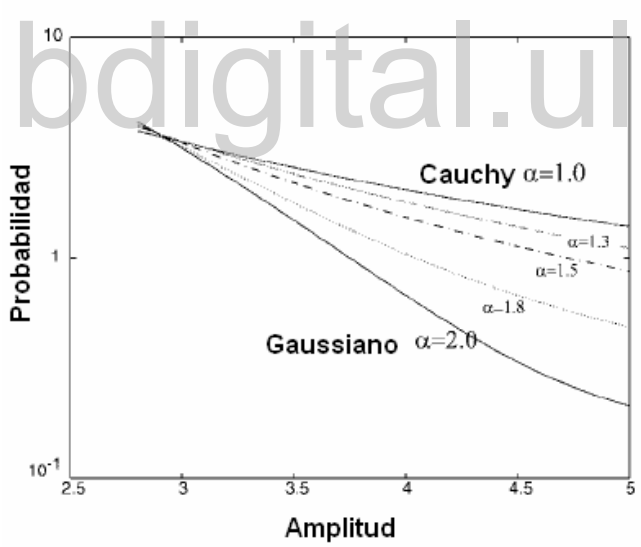

Figura 2.3: Representación de las Colas de la Función de Densidad de la Distribución α-estable para distintos valores de α [13]

Para valores bajos de α, las colas son más pesadas, como se muestra en la figura 2.3 y así el ruido es más impulsivo, mientras que para un valor más grande la distribución

tiene un comportamiento menos impulsivo. En particular, la representación de la distribución Gaussiana viene dada para $\alpha = 2$, y la distribución de Cauchy para $\alpha = 1$, vea la figura 2.3. Para las distribuciones α-estables, los momentos existen solamente para órdenes menores que el exponente característico [14]. El parámetro γ denota la dispersión, la cual determina la extensión de la densidad alrededor de su parámetro de localización y representa el mismo parámetro que la varianza en el modelo Gaussiano.

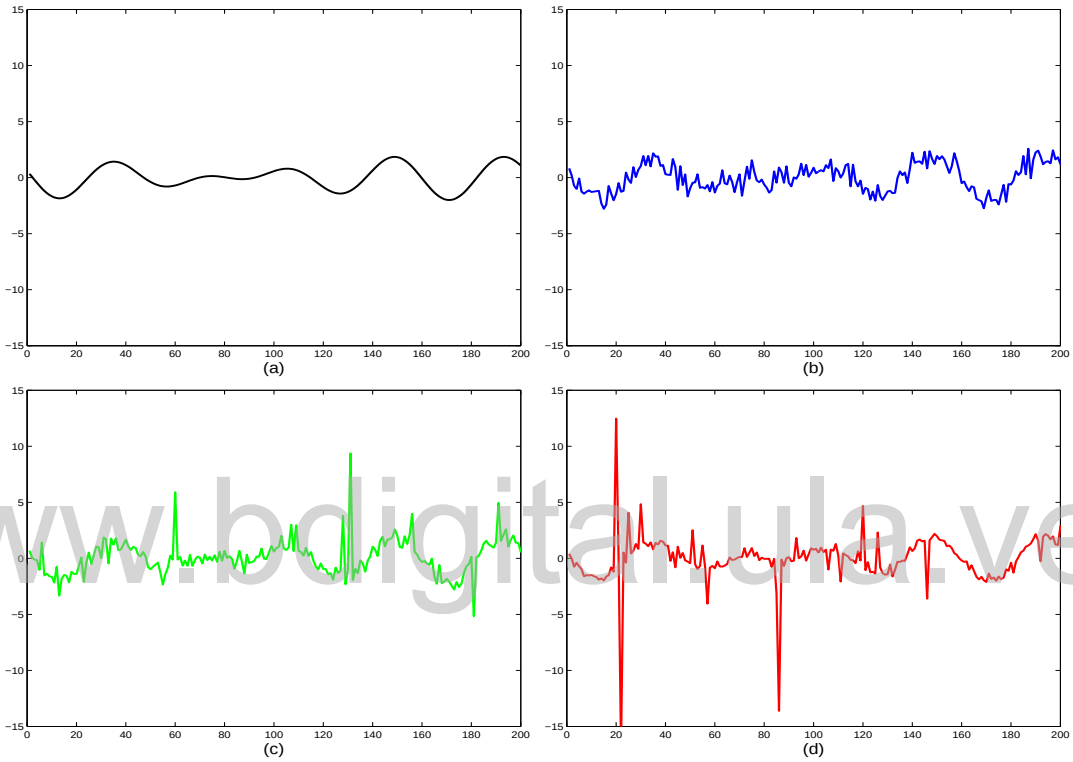

Figura 2.4: Representación de una señal contaminada con ruido tipo α estable. (a) Señal sin ruido. (b) Señal contaminada con ruido Gaussiano ($\alpha = 2$). (c) Señal contaminada con ruido impulsivo intermedio ($\alpha = 1{,}5$). (d) Señal contaminada con ruido Cauchy ($\alpha = 1$)

La impulsividad de la distribución α-estable se puede ver claramente en la figura 2.4. A medida que disminuye el valor de α, son mas frecuentes y de mayor amplitud los impulsos que contaminan la señal. La figura 2.4(a) muestra la señal sin ruido. La ilustración de la misma señal cuando esta contaminada con ruido Gaussiano se muestra en la figura 2.4(b). Con la finalidad de mostrar la impulsividad del ruido modelado con la distribución α-estable se utilizó el parámetro α=1.5, lo que se considera un ruido intermedio, se puede ver en la figura 2.4(c) los impulsos dentro de la señal original.

La figura 2.4(d) muestra la señal contaminada con ruido considerablemente impulsivo para un valor de α=1, cuando la función α-estable toma este valor de α se dice que el ruido generado esta asociado a la distribución de Cauchy.

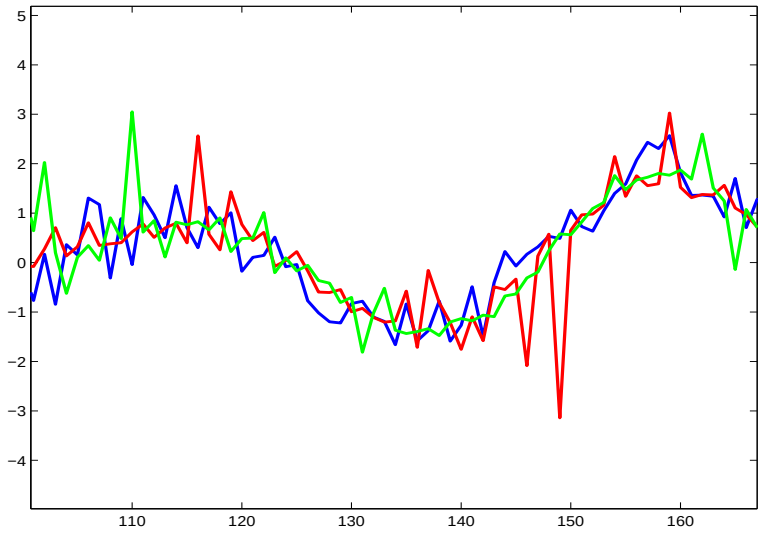

Figura 2.5: Acercamiento de la señal 2.4, donde para distintos valores de α la señal mantiene el mismo patrón

En la figura 2.5 se muestra un acercamiento de las señales generadas por las tres distribuciones distintas mostradas en la figura 2.4, se puede observar que estas señales no presentan muchas diferencia entre ellas, es decir, siguen aproximadamente el mismo patrón. Esta propiedad estimula el uso de distribuciones α-estable, ya que representan situaciones donde el ruido se ha modelado tradicionalmente como Gaussiano, pero en donde los "puntos dispersos"pueden ocurrir.

Capítulo 3

Correlación Tradicional

3.1. Correlación Basada en Momento de Segundo Orden

Es bien conocido que un proceso aleatorio queda completamente caracterizado si se conoce la función de distribución conjunta de las variables aleatorias que conforman el proceso. Definiéndose la función de probabilidad conjunta de n muestras con la siguiente expresión.

$$F(X_1, X_2, ..., X_n) = \Pr[x_1 \leq X_1, x_2 \leq X_2, ..., x_n \leq X_n] \tag{3.1}$$

Sin embargo en muchos casos no se conoce completamente el modelo estadístico del proceso y es suficiente conocer los momentos de primer y segundo orden. Por ejemplo, en el caso de variables aleatorias, el conocer los momentos primer y segundo orden, permiten obtener los parámetros más importantes que caracterizan tales variables, como es el caso de la media, la varianza, la covarianza, la correlación y la densidad espectral de potencia.

Este segmento, se centrará en el cálculo de correlación basada en el momento de segundo orden. Supongamos la variable aleatoria bidimensional (X, Y). El Momento Central M_{ij} de orden $i + j$ del par (X, Y), está definido como [15]:

$$M_{ij} = E[(X - \mu_x)^i (Y - \mu_y)^j] \tag{3.2}$$

donde μ_x y μ_y son las medias de X e Y respectivamente y $E[\cdot]$ denota el valor esperado.

Si se asume que la distribución de los datos del par (X, Y) obedece a la distribución de tipo gaussiano, el momento de segundo orden, cuando $i = j = 1$, se define como:

$$M_{11} = E[(X - \mu_x)(Y - \mu_y)] \tag{3.3}$$

despues de algunas manipulaciones algebraicas la expresión que define el momento de segunod orden se reduce a:

$$M_{11} = E[XY] - \mu_x\mu_y \tag{3.4}$$

conociéndose esta ultima expresión como la covarianza entre X e Y.

La covarianza entre dos variables aleatorias X e Y es una medida de la *relación lineal conjunta* entre dichas variables. Para el caso específico donde las variables X e Y obedecen a la distribución gaussiana con media cero, esta relación queda caracterizada por la $E[XY]$ la cual define el *"Parámetro de Correlación"*, entre ambas variables.

Ahora si se considera que la variable aleatoria (X, Y) es parte de un proceso estocástico, debe tomarse en cuenta su dependencia con el tiempo. Entonces, la correlación entre ambas variables para distintos instantes de tiempo esta representada por:

$$\bar{R}_{XY}(t_1, t_2) = E[X(t_1) \cdot Y(t_2)] \tag{3.5}$$

La ecuación (3.5) es conocida también como correlación cruzada, y permite comparar dos señales diferentes pero coherentes [2]. La función de correlación cruzada contiene información con respecto a las frecuencias que son comunes a ambas señales y a la diferencia de fase entre ellas [2].

Cuando la correlación se calcula entre una señal y una versión desplazada de sí misma se denomina autocorrelación y se define de manera similar:

$$\bar{R}_{XX}(t_1, t_2) = E[X(t_1) \cdot X(t_2)] \tag{3.6}$$

La función de autocorrelación (3.6) permite comparar el grado de similitud entre una señal y una versión desplazada de ella, y se utiliza para medir una banda particular de frecuencias de una señal $X(t)$, así como para detectar una señal repetitiva inmersa en ruido.

3.1.1. Procesos Estacionarios

En las aplicaciones de procesamiento de señales, generalmente se trabaja bajo la suposición que los procesos aleatorios son estacionarios en sentido amplio, así, "las propiedades estadísticas son invariantes respecto a un desplazamiento del origen de los tiempos"[16].

Así , la función de autocorrelación de un proceso aleatorio $X(t)$ está definida como el valor esperado del producto $X(t)\cdot X(t+\tau)$. Suponiendo que el proceso es estacionario en sentido amplio, la autocorrelación sólo dependerá de la diferencia entre las muestras temporales [17], esto es:

$$\bar{R}_{XX}(\tau) = E[X(t) \cdot X(t+\tau)] \tag{3.7}$$

donde el parámetro temporal τ =(t_1 - t_2).

Para todos los procesos estacionarios de interés práctico, se utiliza la llamada hipótesis ergódica como estimador de parámetros a partir de realizaciones particulares (muestras) del proceso dentro de un intervalo finito de tiempo. Se dice que un proceso es ergódico si sus promedios estadísticos coinciden con sus promedios temporales [18][15].

Entonces, para un proceso ergódico podemos definir la "Función de Autocorrelación"de una señal real $X(t)$ de potencia como:

$$\bar{R}_{XX}(\tau) = E[X(t) \cdot X(t+\tau)] = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} X(t) \cdot X(t+\tau) \cdot dt \tag{3.8}$$

Por ejemplo, sean dos procesos $X(t)$ e $Y(t)$ estacionarios de los que se dispone registros de cada uno, es decir se conoce las muestras $[X_1, X_2, ..., X_n]$ y $[Y_1, Y_2, ..., Y_n]$. Se desea calcular la función de correlación cruzada R_{XY} a partir de estos registros, considerando que el proceso es estacionario en sentido amplio y ergódico, la función de correlación cruzada puede ser estimada mediante los promedios.

$$\bar{R}_{XY} = \frac{\sum_{i=1}^{n} X_i \cdot Y_i}{n} \tag{3.9}$$

3.1.2. Propiedades de la Función de Autocorrelación

La función de autocorrelación de procesos estacionarios en sentido amplio, tiene numerosas utilidades además de importantes propiedades, algunas de la cuales presentamos a continuación:

- El valor de la función de autocorrelación en el origen es igual a la energía de la señal, esto es:

$$\bar{R}_{XX}(0) = E[X(t)^2] \tag{3.10}$$

- Para procesos estacionarios, la función de autocorrelación es una función par:

$$\bar{R}_{XX}(\tau) = \bar{R}_{XX}(-\tau) \tag{3.11}$$

- La función de autocorrelación es máxima en el origen [6][2]:

$$\begin{aligned} D(\tau) &= E[|X(t) - X(t+\tau)|^2] \geq 0 \\ E[X(t)^2] + E[X(t+\tau)^2] &\geq 2E[X(t)X(t+\tau)] \\ 2E[X(t)^2] &\geq 2E[X(t)X(t+\tau)] \\ \bar{R}_{XX}(0) &\geq \left|\bar{R}_{XX}(\tau)\right| \end{aligned} \tag{3.12}$$

  para todo valor de τ.

- La autocorrelación de la suma de variables no correlacionadas es la suma de sus autocorrelaciones. Si

$$Z(t) = X(t) + Y(t)$$

  Entonces

$$\begin{aligned} \bar{R}_{ZZ}(\tau) &= \bar{R}_{XX}(\tau) + \bar{R}_{XY}(\tau) + \bar{R}_{YX}(\tau) + \bar{R}_{YY}(\tau) \\ \bar{R}_{ZZ}(\tau) &= \bar{R}_{XX}(\tau) + \bar{R}_{YY}(\tau) \end{aligned} \tag{3.13}$$

  Siempre que $X(t)$ y $Y(t)$ sean incorreladas, es decir $R_{XY}(\tau) = 0$.

3.1.3. Matriz de Correlación

La función de autocorrelación es el estadístico de segundo orden más importante para procesos aleatorios en tiempo discreto, y se representa a menudo en forma de matriz. Por ejemplo, de un proceso $X(n)$, tomamos un vector de tamaño $p+1$

$$X = [x(0), x(1), ..., x(p)]^T$$

el producto punto del vector de la señal está dado por:

$$XX^H = \begin{bmatrix} x(0)x^*(0) & x(0)x(1) & ... & x(0)x^*(p) \\ x(1)x^*(0) & x(1)x^*(1) & ... & x(1)x^*(p) \\ \vdots & \vdots & & \vdots \\ x(p)x^*(0) & x(p)x^*(1) & ... & x(1)x^*(p) \end{bmatrix}$$

donde H representa el valor Hermitico de X y presenta la transpuesta y conjugada del vector original. Esta operación da como resultado una matriz de tamaño $(p+1)\times(p+1)$. Si consideramos que $X(n)$ corresponde a un proceso estacionario en sentido amplio, se toma el valor esperado y usando la simetría Hermitiana, obtenemos los valores de la matriz de autocorrelación:

$$R_{XX} = E\left[XX^H\right] = \begin{bmatrix} R_{XX}(0) & R^*_{XX}(1) & \dots & R^*_{XX}(p) \\ R_{XX}(1) & R_{XX}(0) & \dots & R^*_{XX}(p-1) \\ \vdots & \vdots & & \vdots \\ R_{XX}(p) & R_{XX}(p-1) & \dots & R_{XX}(0) \end{bmatrix}$$

Se puede observar que la matriz de autocorrelación de un proceso estacionario en sentido amplio, tiene un estructura de un matriz Hermitiana, además de tener todos los términos de cada una de las diagonales iguales. Así, R_{XX} cumple es una matriz Toeplitz Hermitiana [6].

3.2. Correlación Basada en el Principio de Verosimilitud Máxima

El método de los momentos es intuitivo y fácil de aplicar, pero generalmente no lleva a los mejores estimadores, pues aunque genera estimadores consistentes, estos a menudo no son muy eficientes, en particular en presencia de ruido de naturaleza impulsiva. En esta sección se estimará la correlación usando el principio de verosimilitud máxima, el cual, a diferencia del método de los momentos, es un método de estimación consistente y genera estimadores insesgados de varianza mínima [12]. Aunque el resultado final para la estimación de correlación en el caso gaussiano es el mismo al encontrado en la sección anterior, la estimación de correlación usando el principio de verosimilitud máxima sentará las bases para introducir dos nuevos métodos de estimación de correlación. Concretamente, cuando la contaminación de fondo sigue un modelo Laplaciano.

En general, la estimación del parámetro de localización asociada al modelo Gaussiano tradicional establecida en el principio de verosimilitud máxima y derivado a partir de muestras independientes e idénticamente distribuidas (i.i.d), sigue una estructura basada en el operador de media (promedio).

Este resultado, sin embargo, se puede generalizar al operador de media ponderada, extendiendo los conceptos de verosimilitud máxima, y asumiendo que la varianza no es igual para todas las muestras. Específicamente, considere el conjunto de muestras independientes $x_1, x_2, ...x_n$, las cuales obedecen la distribución gaussiana, con media β constante, y con varianza distinta para cada una de las muestras σ_i^2.

Como $X_1, X_2, ...X_n$ son variables independientes, la función de densidad conjunta de la muestra es:

$$f(X_1, X_2, ...X_n) = f(X_1) \cdot f(X_2) \cdots f(X_n)$$

$$f(X_1, X_2, ...X_n) = \left\{ \frac{exp\left(\frac{-(X_1-\beta)^2}{2\sigma_1}\right)}{\sigma_1\sqrt{2\pi}} \right\} \left\{ \frac{exp\left(\frac{-(X_2-\beta)^2}{2\sigma_2}\right)}{\sigma_2\sqrt{2\pi}} \right\} ... \left\{ \frac{exp\left(\frac{-(X_n-\beta)^2}{2\sigma_n}\right)}{\sigma_n\sqrt{2\pi}} \right\}$$

entonces

$$f(X_1, X_2, ...X_n) = \frac{1}{\sqrt{(2\pi)^n}} \frac{1}{\prod_{i=1}^{n} \sigma_i} exp\left(-\sum \frac{(X_i - \beta)^2}{2\sigma_i^2}\right) \tag{3.14}$$

Bajo el principio de verosimilitud máxima, se pretende encontrar β que maximice la función de densidad (3.14). Esto equivale a minimizar la función costo (desviación cuadrática de la suma) [1]:

$$G(\beta) = \left(\sum_{i=1}^{n} \frac{(X_i - \beta)^2}{\sigma_i^2}\right) \tag{3.15}$$

Derivamos (3.15) con respecto a β para obtener el estimador

$$\frac{d(G(\beta))}{d\beta} = \left(\sum \frac{(X_i - \beta)^2}{\sigma_i^2}\right) \tag{3.16}$$

$$\frac{d(G(\beta))}{d\beta} = \left(\sum_{i=1}^{n} \frac{X_i}{\sigma_i^2} - \sum_{i=1}^{n} \frac{\beta}{\sigma_i^2}\right)^2 \tag{3.17}$$

Igualando (3.17) a cero obtenemos el estimador de localización basado en el principio de verosimilitud máxima:

$$\bar{\beta} = \frac{\sum W_i \cdot X_i}{\sum W_i} \tag{3.18}$$

donde $W_i = {}^{1}/_{\sigma_i^2} > 0$ representa la ponderación de cada una de las muestras X_i. Observe que el estimador no es más que la *media ponderada*. Si las variables obedecen la misma distribución, es decir $\sigma_1^2 = \sigma_2^2, \ldots, = \sigma_n^2$, entonces el estimador se reduce a $\frac{\sum X_i}{n}$. Una vez obtenido el estimador de verosimilitud máxima del parámetro de localización, se hará la generalización de la estimación del parámetro de correlación bajo el modelo Gaussiano, tomando en cuenta que para estimar la correlación se debe considerar un conjunto bivariable.

Dado el par X y Y de variables aleatorias que obedecen a la distribución gaussiana bivariante con media cero, la función de densidad bivariante viene dada por:

$$f(X, Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2(1-\rho^2)} \cdot \left[\frac{X^2}{\sigma_X^2} - 2\rho\frac{XY}{\sigma_X\sigma_Y} + \frac{Y^2}{\sigma_Y^2}\right]\right) \tag{3.19}$$

Donde $\rho = \frac{E[XY]}{\sigma_X \sigma_Y}$ es el coeficiente de correlación y $E[XY]$ es el parámetro de correlación entre las variables X e Y. Cuando de las variables aleatorias X e Y se dispone de un registro de observaciones $(X_1, X_2, \ldots, X_n)$ y $(Y_1, Y_2, , \ldots, Y_n)$ respectivamente, el parámetro de correlación maximiza la función de densidad (3.19), y es conocido como la correlación muestral.

$$\bar{R}_{XY} = \frac{1}{n} \sum_{i=1}^{n} X_i \cdot Y_i \tag{3.20}$$

Tomando como base el estimador del parámetro de localización (3.18), se puede extrapolar el resultado, considerando la variable Y_i como el equivalente a las ponderaciones W_i. Hacemos notar que esto obliga que Y_i sea estrictamente positiva.

$$\bar{\beta}^{+} = \frac{\sum_{i=1}^{n} Y_i \cdot X_i}{\sum_{i=1}^{n} Y_i} \tag{3.21}$$

Es posible probar que Ec.(3.21) es la solución de

$$\bar{\beta}^{+} = \arg_{\beta} \text{ínf} \sum_{i=1}^{n} Y_i \cdot (X_i - \beta)^2 \tag{3.22}$$

Para obtener un resultado más general, considerando valores negativos de Y_i se procede con la siguiente operación:

$$Y_i = |Y_i| \operatorname{sgn}(Y_i) \tag{3.23}$$

de donde:

$$\operatorname{sgn}(Y_i) = \begin{cases} 1 & \text{si} \quad Y_i \geq 0 \\ -1 & \text{si} \quad Y_i < 0 \end{cases}$$

De modo que la solución de:

$$\bar{\beta} = \arg_{\beta} \text{ínf} \sum |Y_i| \left(\operatorname{sgn}(Y_i) \cdot X_i - \beta\right)^2 \tag{3.24}$$

esta dada por:

$$\bar{\beta} = \frac{\sum |Y_i| \cdot (\text{sgn}(Y_i) \cdot X_i)}{\sum |Y_i|} = \frac{\sum_{i=1}^{n} Y_i \cdot X_i}{\sum_{i=1}^{n} |Y_i|} \tag{3.25}$$

Observe que el signo de la muestra Y_i ha sido desacoplado y pasado a la muestra X_i, además se tiene que el estimador dado en la Ec.(3.18) es proporcional a R_{XY}:

$$\bar{R}_{XY} = \bar{Y}_\alpha \cdot \bar{\beta} \tag{3.26}$$

Donde $\bar{Y}_\alpha = \frac{1}{n} \sum |Y_i|$ es la media de los $|Y_i|$.

Capítulo 4

Correlación Basada En El Operador De Mediana

Así como la estimación de correlación tradicional, se deduce partiendo del modelo gaussiano, la estimación de correlación basada en el operador de mediana tiene sus orígenes en el modelo Laplaciano y presenta una estructura sorprendentemente simple. Distinta de la correlación lineal, la correlación de mediana es robusta en presencia de ruido de naturaleza impulsiva.

4.1. Estimación del Parámetro de Localización

Sea $X = \{X_1, X_2, \ldots, X_n\}$, un conjunto de muestras independientes, distribuidas siguiendo el modelo Laplaciano, donde la media común es β y cada muestra tiene cierta varianza σ_i.

La función de densidad conjunta que describe estadísticamente el comportamiento de estas variables viene dada por:

$$f(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} \frac{1}{\sigma_i} \exp^{-|X_i - \beta|/\sigma_i} \tag{4.1}$$

Usando el principio de verosimilitud máxima se puede encontrar el valor del parámetro de localización (β) que maximice f. Esto es equivalente a minimizar la función costo para procesos Laplacianos, es decir, minimizar la suma de desviaciones absolutas ponderadas:

$$L(\beta) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} |X_i - \beta| \tag{4.2}$$

Partiendo de (4.2) y considerando que $W_i = \frac{1}{\sigma_i^2}$, se quiere conseguir el valor de β tal que se minimice la función:

$$L(\beta) = \sum_{i=1}^{n} W_i \cdot |X_i - \beta| \tag{4.3}$$

El valor $\tilde{\beta}$ que minimiza (4.3) es la mediana ponderada, originalmente introducida hace más de 100 años por Edgeworth [19] definida como:

$$\tilde{\beta} = MED(W_1 \diamond X_1, W_2 \diamond X_2, \cdots, W_n \diamond X_n)$$

donde W_i representa la ponderación de la muestra X_i, y $\diamond$ es el operador réplica definido tal que $W_i \diamond X_i$ significa repetir W_i veces la muestra X_i, siempre que W_i sea entero positivo.

Para el caso en que W_i sea un número entero positivo, la mediana ponderada se calcula como sigue:

1. Reproducir la muestra X_i tantas veces lo indique la ponderación W_i.
2. Ordenar las muestras X_i repetidas en forma creciente .
3. El resultado es la mediana de las muestras repetidas y ordenadas.

A manera de ejemplo se mostrará el calculo de la mediana ponderada del conjunto de muestras $[2, 3, 1, 5, -1]$ con las ponderaciones $[1, 2, 3, 2, 1]$ siguiendo el algoritmo descrito.

El resultado de la tabla 4.1 viene dado por la mediana de las muestras ponderadas y ordenadas, la cual en este caso es 2.

Tabla 4.1: Representación del Cálculo de Mediana Ponderada Para Valores Enteros

Conjunto de Muestras	2		3		1		5		-1
Ponderaciones	1		2		3		2		1
Muestras Ponderadas	2	3	3	1	1	1	5	5	-1
Muestras Ponderadas Ordenadas	-1	1	1	1	$\underline{2}$	3	3	5	5

Para el caso en que W_i sea un número real cualquiera, el operador de réplica $\diamond$, ha sido apropiadamente extendido por Arce en [19]. Considerando que (4.3) admite sólo valores positivos, la idea es modificar la función costo tal que admita valores reales:

$$L(\beta) = \sum_{i=1}^{n} |W_i| \cdot |\text{sgn}(W_i)X_i - \beta| \tag{4.4}$$

Asi para un conjunto de ponderaciones W_i y un conjunto de observaciones X_i, el parámetro de localización estimado es:

$$\tilde{\beta} = MED(|W_i| \diamond \text{sgn}\,|W_i|\,X_i|_{i=1}^{n}) \tag{4.5}$$

$\tilde{\beta}$ recibe el nombre de "*Mediana Ponderada*" y se determina siguiendo los siguientes pasos:

1. Calcule el umbral $T_0 = 1/2 \sum_{i=1}^{n} |W_i|$.
2. Pasar el signo de las ponderaciones a las muestras S_i=sgn$(W_i)X_i$, para $i = 1, 2, \cdots, n$
3. Ordenar las observaciones muestrales "signadas".
4. Sume las magnitudes de las ponderaciones correspondientes a las muestras "signadas" ordenadas, comenzando con el máximo y continuando en forma decreciente.
5. La salida es la muestra "signada" en la cual se cumple que la suma de la magnitud de las ponderaciones es mayor o igual al umbral.

A manera de ilustrar el procedimiento del cálculo de Mediana ponderada para valores reales de W_i, considere un conjunto de muestras dado por [2, 3, -1, 5, -1], y las ponderaciones [-0.1, 0.2, 0.3, -0.2, -0.1]. Si se calcula el umbral siguiendo el paso 1, se tiene que T_0=0.45.

Tabla 4.2: Representación del Cálculo de Mediana Ponderada Para Valores Reales

Conjunto de Muestras	2	2	-1	5	-1
Ponderaciones	-0.1	0.2	0.3	-0.2	-0.1
Muestras Signadas Ordenadas	-5	-2	-1	1	2
Valor Absoluto de las Ponderaciones	0.2	0.1	0.3	0.1	0.2
Suma Parcial de las Ponderaciones	0.9	0.7	<u>0.6</u>	0.3	0.2

Entonces, la mediana ponderada en este caso es -1, ya que al hacer la sumatoria de las ponderaciones de derecha a izquierda la muestra asociada a la primera ponderación que en la sumatoria supera el umbral $T_0 = 0{,}45$ corresponde a -1.

De manera general la mediana ponderada puede definirse como:

$$\tilde{\beta} = \left\{ S_k : \min_k \ tal \quad que \ \sum_{i=0}^{k} \left| W_{l_{(n-i)}} \right| \geq T_0 \right\}$$

donde

$$T_0 = \frac{1}{2} \sum_{i=0}^{n} |W_i|$$

$$S_i = \text{sgn}(W_i) X_i$$

$l_{(k)}$ se refiere a la locación del estadístico de orden k de $S_{(i)}$, donde $S_{(1)} \leq S_{(2)} \leq ... \leq S_{(n)}$ [19].

4.2. Estimación del Parámetro de Correlación

Aunque la correlación muestral tradicional definida en (3.20) es un eficiente estimador, es frágil ante ruido y puntos influyentes. A fin de conseguir un estimador más

robusto, se reformuló el problema de estimación del parámetro de correlación ante procesos gaussianos, utilizando el modelo Laplaciano multivariable descrito por:

$$f(X,Y) = \frac{1}{\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \cdot K_0\left(\sqrt{\frac{2}{(1-\rho^2)}\left[\frac{X^2}{\sigma_1} - \frac{2\rho XY}{\sigma_1\sigma_2} + \frac{Y^2}{\sigma_2}\right]}\right) \tag{4.6}$$

donde $K_0(\cdot)$ es una función modificada de Bessel de orden cero y ρ es el coeficiente de correlación.

Si se intenta aplicar el principio de verosimilitud máxima para determinar el parámetro de correlación (ρ) entre las variables X e Y, se encuentra que tal procedimiento es matemáticamente complejo por lo que se recurre a la analogía entre el operador de mediana y la media [19]. Se asume que si para el proceso gausssiano se cumplen todos los artificios matemáticos, por analogía debería cumplirse en el proceso Laplaciano [1]. Es decir, si en el modelo gaussiano el parámetro de correlación se puede expresar en función del parámetro de localización, tal como indica la expresión (3.26), en el modelo Laplaciano se puede realizar similar por analogía.

Así, considerando que para el caso bivariable no tenemos ponderaciones sino variables, se sustituyen las ponderaciones W_i por la variable Y_i.

$$\tilde{\beta} = \arg_\beta \text{ínf} \sum_{i=1}^{n} |Y_i| \cdot |\text{sgn}(Y_i)X_i - \beta| \tag{4.7}$$

$$\tilde{\beta} = MED(|Y_i| \diamond \text{sgn}\,|Y_i|\, X_i|_{i=1}^{n}) \tag{4.8}$$

Nótese que la expresión para el cálculo de correlación muestral Ec. (3.20) viene dada por el producto del promedio del valor absoluto de las muestras Y_i y la suma ponderadas de las muestras X_i. Además, dado que la media ponderada es el parámetro de localización para el caso gaussiano y que la mediana ponderada corresponde al parámetro de localización para el caso Laplaciano, la Ec.(3.20) puede re-escribirse como:

$$\tilde{R}_{XY} = \bar{Y}_\alpha \cdot \tilde{\beta} \tag{4.9}$$

donde $\tilde{\beta}$ está definido en (4.8) y es el parámetro de localización de las muestras X_i ponderadas por las muestras $|Y_i|$. Nótese que se ha sustituido la media ponderada por

la mediana ponderada. Además, observe que las ponderaciones en este caso no son fijas si no que dependen de las muestras Y_i. Según el valor que tome el parámetro de escalamiento $\bar{Y}_\alpha$, la estimación de correlación varía.

4.2.1. Correlación Mediana Muestral

Si $\bar{Y}_\alpha$ es el promedio de la magnitud de los Y_i, en este caso la expresión (4.9) recibe el nombre de Correlación Mediana Muestral y está dada por:

$$\tilde{R}_{XY} = \left(\frac{1}{n}\sum_{i=1}^{n}|Y_i|\right) \cdot MED(|Y_i| \diamond \operatorname{sgn}(Y_i)X_i|_{i=1}^{n}) \tag{4.10}$$

4.2.2. Covarianza Mediana Muestral

La expresión (4.10) puede ser poco robusta debido a la operación de promedio involucrada en el calculo de correlación, por ejemplo en el caso de que una de las muestras Y_i es un impulso. A fin de dar mayor robustez se sustituye $\bar{Y}_\alpha$ por el parámetro de localización correspondiente al modelo Laplaciano, obtiéndose un estimador más robusto denominado Covarianza Mediana Muestral [1].

$$\widehat{R}_{XY} = MED(|Y_i|\,|_{i=1}^{n}) \cdot MED(|Y_i| \diamond \operatorname{sgn}(Y_i)X_i|_{i=1}^{n}) \tag{4.11}$$

4.2.3. Autocorrelación Mediana Muestral

Los conceptos anteriormente descritos pueden extenderse al caso de que las muestras Y_i provengan del mismo proceso que X_i, es decir, que Y_i sea una versión desplazada en tiempo de X_i. En este caso se habla de la función de autocorrelación y viene dada por:

$$\tilde{R}_{XX} = \left(\frac{1}{n}\sum_{i=1}^{n}|X_i|\right) \cdot MED(|X_i| \diamond |X_i||_{i=1}^{n}) \tag{4.12}$$

Claramente la correlación muestral basada en la mediana, es no simétrica, pues $X_i \diamond Y_i \neq Y_i \diamond X_i$. Para obtener una matriz simétrica de autocorrelación se procede de la siguiente manera: se toman M versiones desplazadas del vector X_i, cada una con

dimensión N. Denotamos por X_i el vector X desplazado i muestras para $i = 1, 2 \cdots M$. Calculamos la correlación basada en la mediana entre los vectores.

$$\tilde{R}_{1,i} = \left(\frac{1}{n} \sum_{j=1}^{n} |X_{i,j}| \right) \cdot MED \left(|X_{i,j}| \diamond \operatorname{sgn}(X_{i,j}) X_{1,j}|_{j=1}^{n} \right) \tag{4.13}$$

para $i = 1, 2...M$ y $j = 1, 2...N$. Donde $X_{i,j}$ es la j-ésima muestra del vector X_i. La ecuación (4.13), define la llamada correlación por adelanto en mediana (*Forward Correlation*). En forma análoga se calcula la correlación por retraso basada en Mediana (*Backward Correlation*) $\tilde{R}_{i,1}$. Luego, se considera el promedio entre ambas correlaciones $\tilde{R}_i = (\tilde{R}_{1,i} + \tilde{R}_{i,1})/2$ donde $i = 1, 2...M - 1$ y se genera una matriz de Toeplitz $\tilde{R}$ a partir de los R_i, dicha matriz $\tilde{R}$ es la matriz de correlación basada en mediana [1].

4.2.4. Autocovarianza Mediana Muestral

Otra posibilidad que es útil para comparar una señal con una versión desplazada de si misma en el tiempo es la Covarianza Mediana definida como:

$$\widehat{R}_{XX} = MED(|X_i|\,|_{i=1}^{n}) \cdot MED \left(|X_{i,j}| \diamond \operatorname{sgn}(X_{i,j}) X_{1,i}|_{i=1}^{n} \right) \tag{4.14}$$

Se construye la matriz de covarianza de mediana $\hat{R}$ (toeplitz) de manera análoga, usando el promedio de la covarianza por adelanto y por retraso.

$$\hat{R}_{1,i} = MED \left(|X_{i,j}||_{j=1}^{n} \right) MED \left(|X_{i,j}| \diamond (X_{i,j})\ X_{1,j}\|_{j=1}^{n} \right) \tag{4.15}$$

donde $i = 1, 2, \cdots, M$ y $j = 1, 2, \cdots, N$.

4.3. Comparación del Método Tradicional y los Métodos Basados en Mediana

A manera de ejemplo, se hará una comparación del método de correlación tradicional expresado en (3.9), con respecto a los métodos robustos basados en mediana propuestos en [1]. Considere el conjunto de muestras $\{4, -3, -1, 1, -1, 3, -1, 0, 4, 1, -3\}$. El resultado generado por el cálculo de correlación, a partir de los conceptos antes descritos fueron para la autocorrelación tradicional $\bar{R}_{XX}(0)$=5.8. Para la autocorrelación

de mediana $\tilde{R}_{XX}(0)$=6, y para la autocovarianza de mediana $\hat{R}_{XX}(0)$=3. Se modificó la primera muestra sustituyendo la muestra $\underline{X_1 = 4}$ por la muestra con valor $X_1 = 10$, la cual se considera un impulso, ya que esta alejado del valor de los datos. Se realizó nuevamente el cálculo de correlación dando como resultado para la autocorrelación tradicional ,autocorrelación de mediana y autocovarianza de mediana $\bar{R}_{XX}(0)$=13.5, $\tilde{R}_{XX}(0)$=10.2, $\hat{R}_{XX}(0)$=4, respectivamente. La correlación tradicional ante la presencia del impulso se aleja notablemente del valor real, en comparación con la correlación y covarianza de mediana, que intentan mantener el valor real de correlación para una señal no contaminada, en este caso la covarianza de mediana funciona mucho mejor que la correlación de mediana.

Capítulo 5

Covarianza con Determinante Mínimo (MCD)

El método Covarianza con Determinante Mínimo (MCD, del ingles *Minimum Covariance Determinant*) es un estimador robusto para el parámetro de localización y la dispersión de datos multivariados. El objetivo es conseguir un conjunto de h observaciones (de un total de n muestras) tal que la matriz de covarianza asociada a dicho subconjunto tenga el menor determinante posible. En el presente trabajo se utilizó el algoritmo MCD-FAST, introducido por Rousseaw et al [4] y disponible en [20], el cual es bastante eficiente en términos de tiempo computacional.

5.1. Principio de la Implementación

En la implementación del algoritmo se asume que el valor del número de observaciones n es por lo menos 5 veces el número de variables p (dimensión de la matriz de correlación). Si esto no ocurriera, entonces habría que reducir de antemano el valor de p bien sea por selección de variables o usando análisis de componentes principales.

El valor de h usado en este trabajo es $\frac{3}{4}n$, pero en realidad se puede considerar cualquier valor h del intervalo $[n/2, n]$. El algoritmo Fast-MCD selecciona (de manera iterativa) las muestras que conformarán el subconjunto final, con ayuda de la denominada "distancia robusta basada en MCD", la cual no es más que una versión modificada

de la distancia de Mahalanobis [21]. A saber, si H es un subconjunto de h observaciones, donde cada observación es un vector de tamaño p, μ_H es la media tradicional de los elementos de H y S_H es la matriz de covarianza tradicional de los elementos de H:

$$\mu_H = \frac{1}{h} \sum_{X_i \in H} X_i \tag{5.1}$$

$$S_H = \frac{1}{h} \sum_{X_i \in H} (X_i - \mu_H)(X_i - \mu_H)^T \tag{5.2}$$

donde h es la cardinalidad del conjunto H y T denota la transpuesta.

La distancia robusta se define para todo X_i del universo de muestras:

$$d_H(X_i) = \sqrt{(X_i - \mu_H)' S_H^{-1} (X_i - \mu_H)} \tag{5.3}$$

El proceso es iterativo en el sentido de que minimizando la distancia robusta, se obtienen nuevos subconjuntos H_k tales que la nueva matriz de covarianza S_k tiene determinante menor ó igual que la matriz S_{k-1} asociada al subconjunto H_{k-1} anterior.

5.2. Justificación de la Distancia Robusta

La intención de calcular la distancia es separar la data en dos partes, una definida por los datos buenos y la otra por los puntos dispersos, tomando solo la señal buena tal que se pueda obtener una mejor estimación del MCD. La efectividad de la distancia robusta se debe al proceso iterativo al que es sometida la distancia de Mahalanobis, el cuál hace que se rechacen más datos que se consideran impulsos, y por lo tanto que no pertenecen a la señal original.

Se ilustrará con un ejemplo el efecto de rechazo de puntos dispersos usando la distancia de Mahalanobis y el efecto de rechazo de puntos dispersos usando la distancia robusta basada en MCD. Considere una señal contaminada con ruido impulsivo, se pretende hacer una comparación entre la distancia robusta basada en MCD, y la distancia de Mahalanobis tradicional. En este ejemplo se considera una señal sinusoidal contaminada con ruido impulsivo donde el número de observaciones es n=220 y el número de variables es p=20.

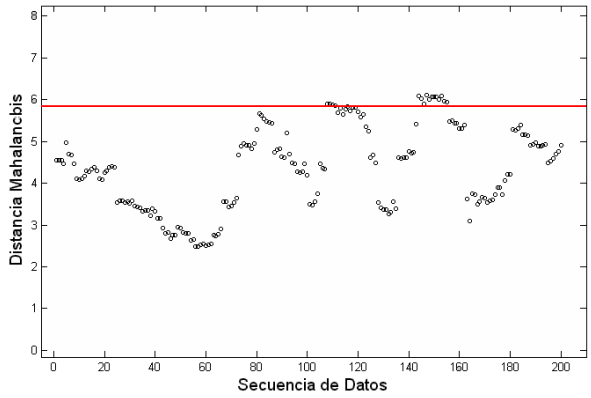

Figura 5.1: Gráfico que muestra la Distancia de Mahalanobis para una señal sinusoidal contaminada con ruido impulsivo.

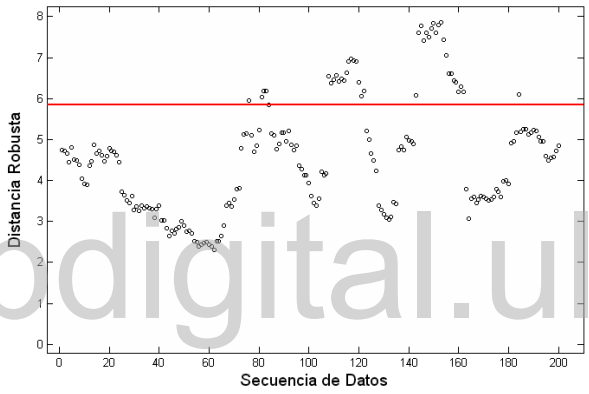

Figura 5.2: Gráfico que muestra la Distancia Robusta Basada en MCD para una señal sinusoidal contaminada con ruido impulsivo.

Observe las figuras 5.1 y 5.2, muestran la distancia de Mahalanobis y la distancia robusta respectivamente. La línea continua en cada uno de los gráficos es el umbral dado por $\sqrt{\chi^2_{20,0,975}}$ el cual indica el margen de separación entre los datos buenos y los puntos dispersos. La Figura 5.1 la cantidad de datos dispersos rechazados por la distancia de Mahalanobis. La figura 5.2, esta en comparación con la figura 5.1 rechaza mayor cantidad de datos relacionados con el ruido impulsivo, lo cúal genera robustez en el momento de hacer el cálculo de MCD.

La figura 5.3 muestra más claramente el efecto de la cantidad de datos impulsivos rechazados por cada una de las distancias, mostrando una comparación entre la

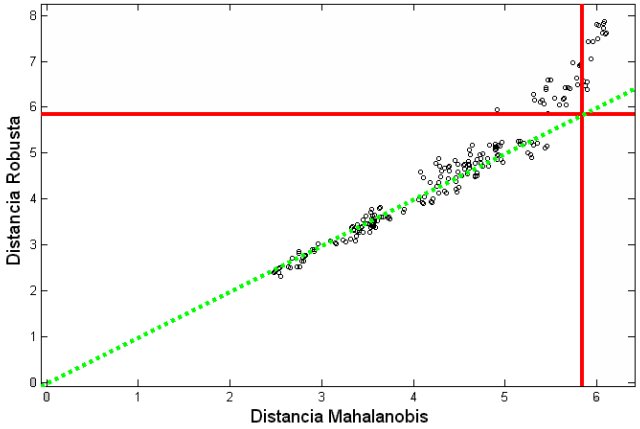

Figura 5.3: Distancia Robusta basada en MCD vs Distancia de Mahalanobis.

distancia de Mahalanobis y la distancia Robusta basada en MCD. La línea continua vertical _____ y la línea continua horizontal _____ que se encuentran dentro del gráfico representan el umbral para la distancia de Mahalanobis y la distancia robusta respectivamente. Basados en la cantidad de datos dispersos rechazados, note que la distancia de Mahalanobis detecta pocos datos dispersos. La distancia robusta basada en MCD detecta una gran cantidad de datos considerados impulsos por lo tanto refleja una mejor estimación de correlación considerando que rechaza más datos dispersos los cuales no pertenecen a la señal original, lo que le genera robustez a dicho método.

5.3. Algoritmo del Proceso MCD

El objetivo es obtener nuevos subconjuntos H_k a partir del conjunto inicial H, partiendo de un proceso iterativo en el cual se mínimiza la distancia robusta. La matriz de covarianza S_k asociada al subconjunto H_k debe tener determinante menor ó igual a la matriz de covarianza S_{k-1} asociada al subconjunto H_{k-1} anterior. Así, para un conjunto de datos $X = \{X_1, X_2, ..., X_n\}$ donde cada X_i tiene p componentes, y $n \leq 600$ el MCD se determina siguiendo los siguientes pasos: (para el caso donde $n > 600$ ver [4])

Paso 1 Para establecer las condiciones iniciales, construya 500 veces un subconjunto H_1 de h observaciones, donde $h < n$. Proceda de la siguiente manera:

- Tome aleatoriamente un subconjunto J de tamaño $(p+1)$, calcule la media $(\mu_0(J))$ y la matriz de covarianza $(S_0(J))$ asociada dicho subconjunto.
- Calcule el determinante de S_0; si $det(S_0) = 0$, entonces aumente en una muestra aleatoria a J , y continúe aumentando muestras hasta que se cumpla que $det(S_0) > 0$.
- Calcule la distancia robusta $d_0(X_i)$ para $i = 1, 2, ...n$.
- Ordene las distancias de manera ascendente, tal que se obtenga una permutación π que cumpla:

$$d_0(X_{\pi(1)}) \leq d_0(X_{\pi(2)}) \leq \cdots \leq d_0(X_{\pi(n)})$$

- Tome los puntos X_π correspondientes a las h menores distancias y forme un nuevo subconjunto H_1.

$$H_1 = \left\{X_{\pi(1)}, X_{\pi(2)}, \cdots, X_{\pi(h)}\right\}$$

Paso 2 Aplique 2 veces el algoritmo del C-paso, descrito a continuación:

- Tome el subconjunto H_k anterior y calcule su media μ_k y su covarianza S_k asociada.
- Calcule las distancias $d_k(X_i)$ para $i = 1, 2 \cdots n$
- Ordene las distancias generando una permutación π, tal que se cumpla:

$$d_k(X_{\pi(1)}) \leq d_k(X_{\pi(2)}) \leq \cdots \leq d_k(X_{\pi(n)})$$

- Tome los puntos X_π correspondientes a las h menores distancias y forme un nuevo subconjunto H_{k+1}.
- Calcule μ_{k+1} y S_{k+1}.

Paso 3 Tome los 10 resultados con menor determinante de S_{k+2} y ejecute el C-paso hasta que converjan.

Paso 4 Reporte la solución μ y S con mas bajo det(S)

Se puede decir que el C-Paso es la base del algoritmo MCD, ya que se tiene el siguiente resultado:

Si H_{k-1} es un subconjunto obtenido a partir de H_k mediante la aplicación de un C-Paso, entonces se tiene que:

$$\det(S_{k-1}) \leq \det(S_k)$$

Es decir, cada vez que se aplica un C-Paso, se obtiene una matriz de covarianza con determinante menor o igual al anterior. Considerando que el número de posibles subconjuntos de H_k que pueden obtenerse es finito, y que la sucesión de determinante,

$$\{\det(S_k)\}_k$$

es una sucesión no creciente acotada inferiormente, se puede garantizar que dicha sucesión converge a un valor mínimo. Los pasos (3) y (4) descritos anteriormente se encargan de tomar las "mejores"sucesiones $\{\det(S_k)\}_k$ y escoger el punto de convergencia menor. Cabe notar que el hecho de tener una matriz de covarianza S_k que posea un determinante pequeño, indica que el conjunto H_k (asociado$< S_k$) contiene datos que estan poco dispersos.

5.4. Normalización del Proceso MCD

Para concluir el algoritmo es necesario hacer algunas normalizaciones. Si H_f es el subconjunto final, tal que la matriz de covarianza asociada a S_f tiene determinante mínimo, entonces el parámetro de localización esta dado por $\mu_{MCD} = \mu_f$ y la dispersión por la matriz:

$$S_{MCD} = \frac{MED_{\overrightarrow{X}_i}\left(d^2_{(\mu_f,S_f)}(\overrightarrow{X_i})\right)}{\chi^2_{p,1/2}} S_f \tag{5.4}$$

donde $d^2_{(\mu_f,S_f)}(\overrightarrow{X_i})$ es la distancia robusta asociada al subconjunto Hf, y $\chi^2_{p,1/2}$ es el percentil $1/2$ de la distribución Chi Cuadrado p-dimensional.

Este factor se introduce para obtener consistencia en el caso en que los datos provengan de una distribución normal.

Como último paso, se hace una re-ponderación dada por:

$$\mu = \frac{\left(\sum_{i=1}^{n} w_i \overrightarrow{x}_i\right)}{\left(\sum_{i=1}^{n} w_i\right)} \tag{5.5}$$

$$S = \frac{\left(\sum_{i=1}^{n} w_i \left(\overrightarrow{X}_i - \overrightarrow{\mu}\right)\left(\overrightarrow{X}_i - \overrightarrow{\mu}\right)'\right)}{\left(\sum_{i=1}^{n} w_i - 1\right)} \tag{5.6}$$

donde

$$w_i = \begin{cases} 1 & \text{si} \quad d_{(T_{MCD}, S_{MCD})}(i) \leq \sqrt{\chi^2_{p,0,975}} \\ 0 & \text{en} \quad \text{otro} \quad \text{caso} \end{cases}$$

La distancia robusta obtenida con este método permite detectar fácilmente muestras causadas por ruido impulsivo, debido a que ellas tienen una mayor distancia de Mahalanobis. Este método funciona para casos tan extremos como en el que el 49 % de las muestras se consideran impulsos y por lo tanto no son confiables.

5.5. Comparación del Método Tradicional y el Método Basado en MCD

A manera de ejemplo, se hará una comparación del método de correlación tradicional expresado en (3.9), con respecto a el método robusto de correlación basado en MCD propuesto en [4].

Considere el sistema de muestras $\{4, -3, -1, 1, -1, 3, -1, 0, 4, 1, -3\}$. Se pretende hacer una comparación del cálculo de correlación basado en el método tradicional, con el método basado en MCD. El resultado generado por la autocorrelación tradicional es $\bar{R}_{XX}(0)$=5.8, y por la correlación basada en MCD $\hat{S}(0)$=6.6. Se modificó la primera muestra sustituyendo la muestra $\underline{X_1{=}4}$ por la muestra con valor $X_1 = 10$, la cual se puede considerar un impulso que no tiene nada que ver con la secuencia de datos. Se realizó nuevamente el cálculo de correlación dando como resúltado para la autocorrelación tradicional , y para la correlación basada en MCD $\bar{R}_{XX} = 13{,}5$, $\hat{S} = 9{,}8$,

respectivamente. Cuando se cálcula el valor de $\hat{S}$ en presencia del impulso, este en comparación con $\bar{R}_{XX}$ trata de mantenerse cerca del valor original de correlación. La correlación tradicional ante impulsos se aleja más del valor real que la la correlación basada en MCD, resultado que muestra la robustez del método propuesto.

Capítulo 6

Aplicaciones de Prueba

A continuación, se muestra un conjunto de aplicaciones de carácter práctico en el que los tres métodos de estimación de correlación discutidos en los capítulos anteriores son utilizados.

6.1. Descripción de la Longitud Mínima (MDL)

La mayoría de señales que se encuentran en la naturaleza son generalmente repetitivas a través del tiempo, y los puntos colindantes están altamente correlacionados. En muchos problemas prácticos, la señal de interés puede ser modelada como la superposición de un número finito de señales sinusoidales inmersas en ruido, como por ejemplo el restablecimiento de los polos de un sistema desde su respuesta natural. La clave de estos problemas está en la estimación del número de señales sinusoidales. Generalmente en estos problemas la representación de los datos es muy pobre, y se puede encontrar una mejor interpretación de la señal a través de la compresión de los datos. La intención es separar la información de la señal de ruido que está determinada por la matriz de correlación.

El análisis detrás del principio de la Descripción de la Longitud Mínima (MDL) introducido por Rissanen [8], ofrece una alternativa viable donde el ruido está definido como la parte incompresible de la señal, y consiste en descubrir regularidades en un conjunto de datos; el éxito de encontrar tales regularidades se puede medir por la

longitud con la cual los datos pueden ser descritos [5].

Un enfoque a este problema es considerar que el número de señales puede ser determinado por los autovalores de la matriz de correlación del vector de observación, específicamente por el valor para el cual la función de MDL es mínima. La idea básicamente es encontrar el valor de k que minimice:

$$MDL(k) = -\log\left(\frac{\prod_{i=k+1}^{q}\lambda_i^{1/(q-k)}}{\frac{1}{q-k}\sum_{i=k+1}^{q}\lambda_i}\right)^{(q-k)N} + \frac{1}{2}k(2q-k)\log N \tag{6.1}$$

donde λ_i representa a los autovalores de la matriz de correlación, q es el tamaño del vector de observación, N es el tamaño de la muestra. Así, el número de señales complejas será determinado por el valor $k \in [0, 1, 2, ldots, q-1]$ para el cual (6.1) es mínimo.

En la Ec. (6.1) se puede notar que el criterio del MDL depende de los autovalores de la matriz de correlación, y con la finalidad de comparar los métodos de estimación de correlación descritos en la sección anterior, se sustituirán en el algoritmo los autovalores de la matriz de correlación tradicional impuesta por los autovalores de las matrices de correlación robustas propuestas en [1] y [4].

Por ejemplo, el vector de observación se considera un problema importante en el procesamiento discreto de señales. Considere el vector de observación representado por la señal:

$$X(n) = \sum_{i=1}^{p} A_i \cos(w_i n + \phi_i) + \eta(n) \tag{6.2}$$

Donde A_i son las amplitudes, w_i las frecuencias desconocidas y las fases ϕ_i son asumidas variables independientes distribuidas uniformemente entre $[0, 2\pi)$. Se asume que el ruido representado por el vector $\eta(n)$ es independiente de la señal y es un proceso de tipo α-estable con parámetro de localización cero y características de α desconocidas [22].

El gran problema asociado con el modelo descrito en (6.2), es la estimación del número de señales p, a partir de una serie finita de observaciones $X(n_1)$,$X(n_2)$,...,$X(n_n)$,

donde n es el número de observaciones de dimensión p.

Una solución prometedora para este problema está basada en la estructura de la matriz de correlación del vector de observación $X(.)$. El número de señales p puede ser determinado a partir de los autovalores mas pequeños de la matriz de correlación R, pero el problema es que esta aún no se conoce en la práctica [5]. Generalmente cuando se estima la matriz de correlación a partir de una muestra de tamaño finito, todos los autovalores son diferentes con probabilidad uno, de modo que se hace mas difícil determinar el número de señales, tan solo tomando en cuenta los autovalores. Rissanen, basado en los argumentos del criterio de teoría de información , propuso el modelo de la Descripción de la Longitud Mínima, el cuál será ilustrado en el capítulo 7 con ejemplos mas específicos.

6.2. Clasificación de Múltiples Señales(MUSIC)

Un problema clásico en el procesamiento de señales es la estimación de las componentes de frecuencia de una señal. En la literatura pueden encontrarse técnicas avanzadas como Clasificación de Múltiples Señales (MUSIC, del inglés *Multiple Signal Classification*) para estimar estas componentes. El funcionamiento eficiente de esta técnica de estimación en algunas situaciones, depende de una estimación eficiente de la matriz de correlación.

Un modelado inexacto de la matriz de correlación produce errores en la estimación de componentes de frecuencia obtenidas, esto quiere decir que para obtener un buen desempeño el método MUSIC depende notablemente de un buen cálculo de la matriz de correlación.

El algoritmo MUSIC es un método de alta resolución que genera estimaciones de las componentes de frecuencia de una señal realizando la autodescomposición de la matriz de autocorrelación de un vector de datos.

Este método satisface la estimación del espectro de señales sinusoidales, específicamente para sinusoides inmersas en ruido, y los cocientes de la señal de interferencia son bajos. La clave del algoritmo MUSIC es la forma en que están modelados los datos, por ejemplo, suponga:

$$X(t) = \sum_{i=1}^{p} A_i \cos(w_i t + \phi_i) + \eta(t) \tag{6.3}$$

donde A_i son las amplitudes reales, y las fases ϕ_i son asumidas variables independientes distribuidas uniformemente entre $[0, 2\pi)$. El vector $\eta(t)$ es ruido y se asume sea un proceso tipo α-estable con parámetro de localización cero y características de α desconocidas. Las frecuencias desconocidas están representadas por w_i, y serán estimadas bajo el algoritmo MUSIC. A fin de simplificar el problema, el número de sinusoidales complejas (p) se conoce ó puede ser estimado usando el método de MDL descrito anteriormente en la sección 6.1.

El problema consiste en descubrir cuales son las componentes de frecuencia de una señal inmersa en ruido impulsivo. El algoritmo que describe el proceso para obtener las componentes frecuenciales de la señal inmersa en ruido se muestra a continuación [6]:

1. Calcule la matriz de autocorrelación $\tilde{R}$ de la señal original.

2. Desarrolle la autodescomposición de la matriz $\tilde{R}$:

$$\tilde{R}V = V\Lambda \tag{6.4}$$

donde $\Lambda = diag(\lambda_0, \lambda_1, ..., \lambda_M - 1)$, $\lambda_0 \geq \lambda_1 \geq .... \geq \lambda_{M-1}$ son los autovalores de $\tilde{R}$ y $V = [q_0, q_1, ...q_{M-1}]$ son sus correspondientes autovectores.

3. Estime el número $2\tilde{p}$ de señales complejas, usando el criterio de MDL.

4. Calcule el spectro de MUSIC definido como:

$$\hat{P}(w) = \frac{e^H(w)e(w)}{e^H(w)V_n V_n^H e(w)} \tag{6.5}$$

donde

$$e^H(w) = [\ 1 \quad \cos(w) \quad .... \quad \cos((M-1)w)\ ]$$

$$V_n = [\ q_{2\hat{p}} \quad q_{2\hat{p}+1} \quad ..... \quad q_{M-1}\ ]$$

5. Encuentre los $\hat{p}$ picos más grandes de $\hat{P}(w)$, los puntos donde ocurren dichos picos corresponderán a las componentes de frecuencia de la señal en estudio.

Puede observarse que MUSIC se basa en el cálculo de la matriz de correlación, representando así un ejemplo práctico donde los métodos de correlación estudiados en los capítulos anteriores pueden someterse a prueba a fin de comparar su desempeño.

6.3. Normalización de Datos de Microarray de ADNc

6.3.1. El Experimento de Microarray de ADNc

El microarray de Ácido Desoxirribonucleico complementario (ADNc) es uno de los principales ensayos que evalúan la expresión genética de miles de genes simultáneamente. Esta tecnología consiste en un robot de "pequeñas pipetas" (llamadas print-tip) que incuban el Ácido Ribonucleico mensajero (ARNm) de las muestras a analizar sobre una lámina de arreglo genético. Las láminas son arreglos de pequeños envases llamados *spots* que contienen secuencias conocidas de ADNc [7].

El ensayo de microarray de ADN se inicia con la fabricación de un arreglo de material genético. Este arreglo consiste en un gran número de moléculas de ADNc ordenadas en miles de spots de manera que formen una matriz bidimensional de material genético.

El material genético inmovilizado generalmente es llamado sonda. Las sondas provienen de clones de ADN previamente registrados en librerías génicas. En la figura 6.1 se muestra una ilustración del arreglo de sondas para el ensayo de microarray:

El marcaje es un procedimiento de resaltar las muestras a analizar. Existen varios métodos de etiquetado pero la técnica más usada en el ensayo de microarray de ADNc es el marcaje con fluorocromo. Después del etiquetado, las muestras marcadas se colocan en una matriz de pequeñas pipetas que es controlada por un robot de alta precisión. Este robot incuba las muestras marcadas dentro de los *spots* del arreglo sondas.

Luego de la hibridación entre las sondas y las muestras a analizar, se efectúa la obtención de los datos finales. Inicialmente se obtiene una imagen, pasando el arreglo a través de un escáner que excita los tintes aplicados a las muestras. El escáner primero

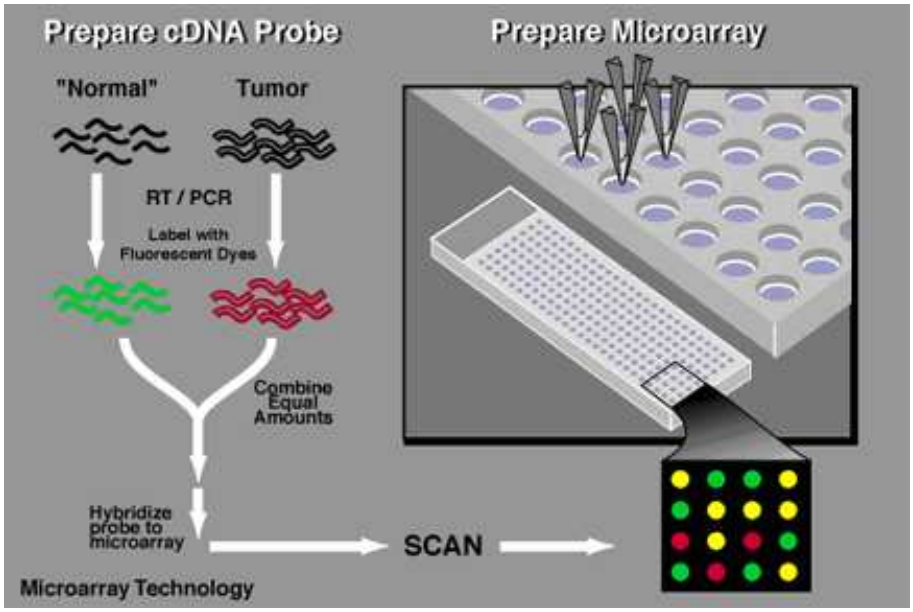

Figura 6.1: Arreglo de sondas para el experimento de Microarray

aplica un láser con una longitud de onda que excita los tintes verdes y luego con una longitud de onda del laser que excita el arreglo estimulando los tintes rojos.

Los resultados de la aplicación del escáner sobre el arreglo genera dos imágenes. Cada imagen es una componente (una en verde y una en rojo) de la imagen definitiva. El resultado final es una imagen en formato TIFF que tiene la capacidad de desplegar la información de 43.000 spots a $2^{16} = 65536$ niveles de intensidades distintos. En la figura 6.2 se muestra un resumen gráfico del experimento completo de microarray.

A partir de las imágenes se obtienen los niveles de expresión genética de cada uno de los spots de la siguiente forma: Para la imagen de cada spot se obtiene un vector con los valores de los píxeles pertenecientes al spot y un vector con los valores de los píxeles del background local. Luego se implementa la operación de mediana al vector que contiene los píxeles del spot y la mediana al vector con los píxeles del background. Finalmente, la expresión genética se obtiene como la mediana del vector de spots menos la mediana del vector del background.

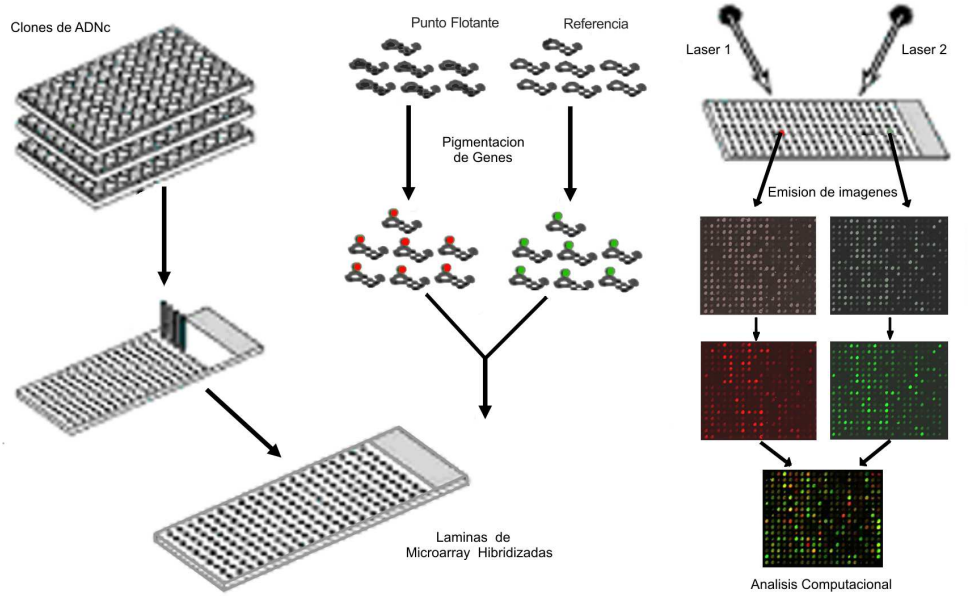

Figura 6.2: Procedimiento del Experimento Microarray de ADNc

6.3.2. Normalización en Microarray de ADNc

En ocasiones las intensidades de rojo y de verde de la imagen son resultados de procesos no dependientes de la interacción biológica de los genes, sino de errores propios del proceso de adquisición. Estos datos erróneos son consecuencia principalmente de la aplicación de la tecnología en el ensayo del microarray. Son varios los factores que impiden obtener datos de forma óptima en una imagen de microarreglo pero los más destacados se pueden clasificar en:

1. Diferencias en las cantidades de tinte vertidas a la muestras.
2. Diferencias en las eficiencia del etiquetado.
3. Mal funcionamiento del escáner.
4. Diferencias en la potencia de los dos láser en el momento de obtención de las componentes de la imagen.

La normalización en microarray de ADNc tiene como objetivo identificar y remover las fuentes de error en las mediciones de la intensidad de fluorescencia. Este tipo de operación se divide en dos objetivos específicos:

1. Eliminar la variación debido a errores en los datos.

2. Mantener las variaciones debido a la interacción biológica de las muestras.

6.3.3. Normalización por Correlación

Estos métodos son aplicados a los resultados derivados de múltiples repeticiones del mismo experimento de microarray de ADNc.

En los métodos de normalización de múltiples láminas se elige una lámina de referencia con sus respectivos genes de control y láminas de punto flotante con sus genes de control ubicados en la misma posición que los situados en la lámina de referencia. La idea fundamental es normalizar los genes de punto flotante con respecto alos genes de referencia mediante una operación de escalamiento. En la Figura 6.3 se muestra una ilustración de lo explicado anteriormente:

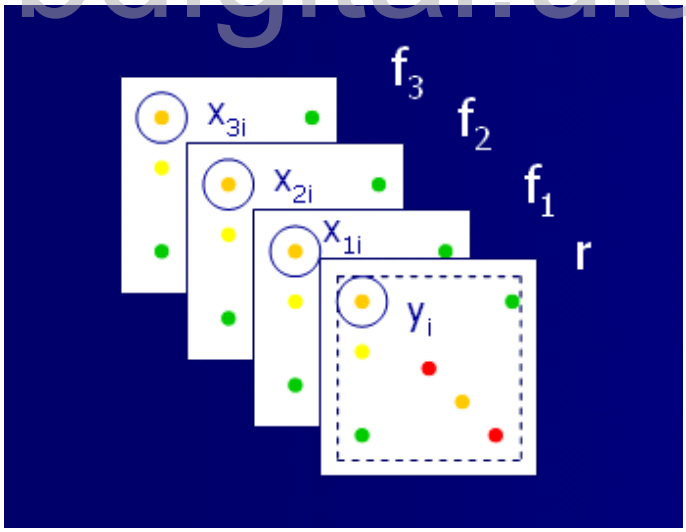

Figura 6.3: Representación de Múltiples Laminas, donde r es le Microarray de Referencia (yi), y f_i son los microarray de punto flotante

La normalización entre múltiples componente busca a través de la operación de

escalamiento los valores más apropiados de a que satisfacen la siguiente expresión:

$$y_i = ax_i \tag{6.6}$$

en donde los y_i son los niveles de expresión genética de los spots en el microarray de referencia y los x_i son los niveles de expresión genética de los spots de los microarray de punto flotante.

Una de las formas de estimar el parámetro de escalamiento a es por medio de la autocorrelación. Este método también es llamado regresión lineal a través del origen, y la gráfica obtenida por el par de conjuntos de datos $\{X_i, Y_i\}|_{i=1}^{n}$ después de la normalización busca una recta de pendiente igual a uno [7].

$$a = \frac{\sum_{i=1}^{N} X_i * Y_i}{\sum_{i=1}^{n} X_i^2} \tag{6.7}$$

Note que la expresión para el cálculo del parámetro de normalización (6.7) depende del cálculo de la matriz de correlación, donde el numerador esta dado por la correlación cruzada entre el microarray de referencia y los microarray control, y el denominador representa la autocorrelación de los microarray de punto flotante.

Capítulo 7

Simulación y Resúltados

Los algoritmos descritos en la sección anterior presentan una dependencia del cálculo de la matriz de correlación, una mala estimación en esta puede originar un resultado equivocado en el cálculo del MDL, MUSIC y la Normalización de Datos de Array de ADNc y en cualquier otro algoritmo que presente dependencia con la matriz de correlación. Con la finalidad de ilustrar el desempeño de los métodos de correlación descritos, se trabajó con simulaciones, sustituyendo en las aplicaciones la matriz de correlación tradicional, por las nuevas matrices robustas propuestas en [1] y [4].

7.1. Estimación de Número de Señales (MDL)

Para ilustrar el desempeño de la matriz de correlación bajo el criterio MDL, considere el problema de modelar una señal como la superposición de p señales sinusoidales de frecuencia y amplitud desconocida. El objetivo es determinar el número de componentes sinusoidales que la señal contiene. La señal de prueba está dada por:

$$X(n) = \sum_{i=1}^{p} A_i \cos(w_i n + \phi_i) + \eta(n) \tag{7.1}$$

En este caso se consideró $p = 3$ y frecuencias de $w_1 = \frac{7}{9}\pi$, $w_2 = \frac{25}{18}\pi$ y $w_3 = \frac{3}{2}\pi$; los A_i son amplitudes reales y las fases ϕ_i se asumen variables aleatorias uniformemente distribuidas en $[0, 2\pi)$. $\eta(n)$ es ruido de naturaleza impulsiva, en este caso un ruido de tipo α-estable [22]. Se tomó un vector $X(t)$ de 220 muestras para generar una matriz

de correlación de tamaño 20×20.

Como en la práctica la matriz de Toeplitz resultante no garantiza estar definida positiva, para atenuar este efecto, algunos de los autovalores más pequeños son omitidos antes de estimar el número de señales usando el criterio MDL.

Con la finalidad de demostrar la robustez de los métodos de correlación propuestos en [1][4], se sustituyen en (6.1) los autovalores de la matriz de Correlación Tradicional impuesta en el método conocido, por los autovalores de las matrices de correlación robustas descritas en los capítulos 4 y 5.

Como un primer ejemplo, se consideró la señal descrita en Ec. (7.1) y se tabularon los resultados generados por el criterio de MDL (6.1), variando los valores de k desde 0 hasta $q - l - 1$, recordando que q representa el tamaño del vector de observación, y donde l es el número de los autovalores más pequeños los cuales serán omitidos, en este caso se consideró $l = 5$.

Analizando los resultados expuestos en la tabla 7.1 se observa que el número de señales $\hat{p}$ estimado a partir del valor de k para el cual el criterio MDL se hace mínimo varía según el método de correlación con el cual fue estimado. Con los métodos de correlación basados en la mediana existe una mejor estimación, como era de esperar, considerando que la señal está contaminada con ruido impulsivo, y el cálculo de los métodos de correlación basado en mediana está modelado bajo esas condiciones. De igual manera el método de Correlación Basado en MCD estimó de manera eficiente el número de componentes de la señal de prueba. Los tres métodos robustos descritos rindieron un valor mínimo para la ecuación (6.1) cuando el valor de $k = 6$; recordando que $k = 2\hat{p}$, el número de señales estimadas es 3. En cambio, observe el valor de k que minimiza el criterio MDL según el método de correlación tradicional, note que la Ec. (6.1) alcanza el valor mínimo cuando $k = 2$, lo cual indicaría que la señal tiene una sola componente de frecuencia, ignorando por completo la existencia de las dos componentes de frecuencia restantes. Este resultado incorrecto se debe a la presencia de impulsos en la señal.

Ahora, con la finalidad de ilustrar el efecto de la impulsividad del ruido en la estimación del número de señales bajo el criterio del MDL, se varió α el cual es un parámetro del ruido tipo α-estable que indica el grado de impulsividad presente. Cuando $\alpha = 2$ se

Tabla 7.1: Valores tabulados del criterio MDL, usando Correlación Tradicional $\bar{R}$,Autocorrelación Mediana Muestral $\tilde{R}$, Autocovarianza Mediana Muestral $\hat{R}$, MCD $\hat{S}$

k	MDL $\bar{R}$	MDL $\tilde{R}$	MDL $\hat{R}$	MCD $\hat{S}$
0	0.6116	∞	∞	∞
1	0.5493	∞	∞	∞
2	**0.4462**	0.8170	0.8099	0.7562
3	0.4631	0.7313	0.6971	0.5866
4	0.4647	0.5895	0.5420	0.5121
5	0.5158	0.5090	0.5126	0.4585
6	0.5695	**0.4431**	**0.4986**	**0.4291**
7	0.6360	0.4726	0.5166	0.4538
8	0.6977	0.4957	0.5237	0.4846
9	0.7554	0.5239	0.5517	0.5138
10	0.8070	0.5467	0.5739	0.5388
11	0.8555	0.5609	0.5906	0.5599
12	0.8987	0.5766	0.5992	0.5759
13	0.9359	0.5855	0.5898	0.5879
14	0.9684	0.5934	0.5934	0.5434

tiene un ruido gaussiano, y a medida que α disminuye ($\alpha > 0$) se tiene un ruido cada vez más impulsivo. El valor de α se varió desde 0.4 hasta 2 para comparar la eficiencia de los distintos métodos de correlación en la estimación del número de componentes ante la impulsividad del ruido . Para hacer esta prueba se utilizó el criterio del MDL, se sustituyó en la Ec. (6.1) los autovalores generados por la matriz de Correlación Tradicional, por los autovalores generados por las matrices de correlación obtenidas mediante los distintos métodos robustos. Se hicieron 500 realizaciones independientes para cada valor de α y se promediaron los resultados obtenidos.

En la Figura 7.1 se muestra el número de señales estimadas bajo el criterio de la "Descripción de la Mínima Longitud", para distintos valores de α. La línea continua _____ describe la estimación del número de señales que genera el MDL calculado con

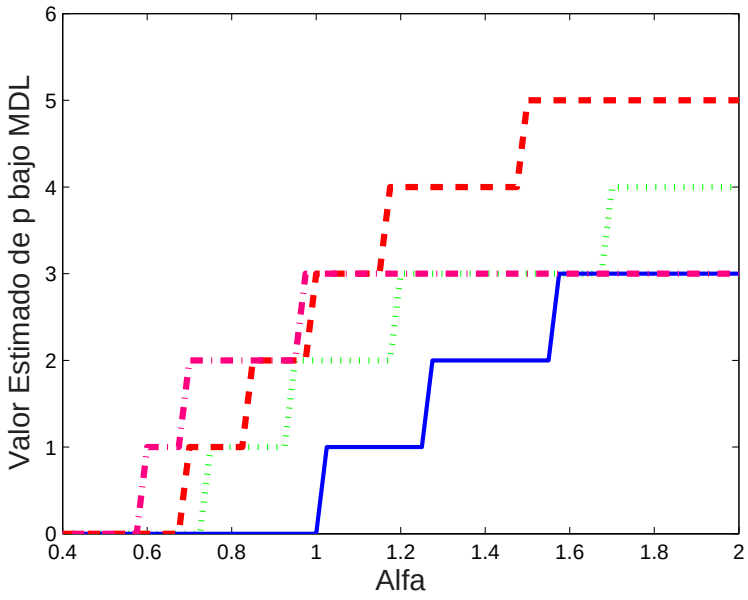

Figura 7.1: Relación entre el parámetro α y el número de señales calculadas bajo MDL, usando ____ para el método de correlación Tradicional,. . . . . el método de Correlación Basada en Mediana, – – – – – el método de Covarianza Basada en Mediana,–. – . – . – .– Correlación Basada en MCD

la matriz de correlación tradicional:observe que mientras el valor de α tiende a 2 la estimación del número de señales se acerca a 3, que es el valor conocido que se desea estimar, lo cual es lógico, ya que α =2 representa un ruido de tipo gaussiano, y se esperaría que la matriz de correlación tradicional funcione eficientemente para estos casos, considerando que para el modelo gaussiano la matriz de correlación tradicional representa el momento de segundo orden. Por lo que la estimación de la matriz de correlación por el método tradicional es optima bajo la suposición que la contaminación de fondo es de tipo gaussiano. La línea punteada $\cdots$ muestra la relación entre los distintos valores de α y la estimación de la señal bajo el criterio del MDL calculado con la matriz de Autocorrelación Mediana Muestral. Se sustituyó en la Ec. (6.1) los autovalores originales del algoritmo por los autovalores generados por la matriz de Autocorrelación Mediana Muestral utilizando la Ec.(4.13). Note en el gráfico que cuando α se encuentra dentro del rango de 1,2 hasta 1,7 el número estimado es $p = 3$; tomando en cuenta que la representación de un ruido impulsivo intermedio puede estar dado por un proceso α-estable con un coeficiente $\alpha \simeq 1{,}5$, se demuestra la robustez

de la Correlación Mediana Muestral ante tal caso. La línea a trozos − − −−, representa el comportamiento del MDL calculado con la matriz de Autocovarianza Mediana Muestral. Se sustituyó en la Ec. (6.1) los autovalores originales del algoritmo por los autovalores generados por la matriz de Autocovarianza Mediana Muestral utilizando la Ec.(4.14). El gráfico muestra que el algoritmo MDL utilizando Autocovarianza Mediana Muestral es robusto ante ruidos considerablemente impulsivos, incluso con valores de α cercanos a 1. Por último, la línea $-.-.-.$, representa el resultado de estimación de número de señales bajo el criterio del MDL utilizando el método de Correlación Basado en MCD. Se sustituyó en la Ec. (6.1) los autovalores algoritmo tradicional del MDL por los autovalores generados por la matriz de Correlación Basada en MCD. Observe en el gráfico que a partir de un valor de $\alpha = 1$(ruido tipo Cauchy) hasta $\alpha = 2$ (ruido tipo Gaussiano) la estimación de número de señales bajo el criterio del MDL es constante en 3, que es el número de componentes de frecuencia contenido en la señal de prueba, el cual se quiere estimar. La extensión del rango de α para el cual se estima eficientemente el número de componentes de la señal muestra la robustez del método propuesto por Rousseauw ante cambios de impulsividad, tanto para los casos en que la señal esta contaminada con ruido de tipo gaussiano, como para los casos en que la señal esta contaminada con ruido impulsivo.

Haciendo un análisis de los resultados en la aplicación del criterio del MDL, se puede decir que el método de Correlación Tradicional funciona perfectamente cuando el ruido es de naturaleza gaussiana, pero falla ante un cierto grado de impulsividad, dando un resultado equivocado del número de señales contenidas en la señal de prueba. Cuando se analiza el gráfico el algoritmo del MDL usando el método de Correlación Basada en Mediana, puede notarse que dio una estimación casi perfecta en presencia de cierta impulsividad, en comparación con el algoritmo del MDL usando la Covarianza Basada Mediana que estimó perfectamente en un rango muy pequeño pero bastante impulsivo; sorprendentemente ambos métodos cuando el ruido inmerso en la señal tenía un comportamiento de tipo gaussiano fallaron, lo que nos hace pensar que son robustos en un cierto rango de impulsividad, correspondiente al caso en que el ruido es de naturaleza Laplaciana, que de hecho es la base a partir de la cual fueron desarrollados dichos métodos. El criterio del MDL utilizando el método de Correlación Basado en

MCD, es el método que estimó el número de componentes de señales de manera más eficiente en comparación con los otros métodos de correlación ante distintos grados de ruido; se supone que es debido a su rechazo a una gran cantidad de puntos dispersos, sin embargo si se considera el tiempo de ejecución de dicho algoritmo se puede decir que es un algoritmo poco eficiente a nivel computacional.

7.2. Clasificación de Señales Multiples (MUSIC)

A continuación, se presenta la simulación de los resultados que ilustra el funcionamiento de los métodos de correlación descritos bajo el algoritmo MUSIC. El problema consiste en descubrir cuales son las componentes de frecuencia de una señal inmersa en ruido impulsivo. El modelo de la señal es:

$$X(t) = \sum_{i=1}^{p} A_i \cos(w_i t + \phi_i) + \eta(t) \tag{7.2}$$

Donde A_i son las amplitudes reales, la fase ϕ_i es asumida un variable independiente distribuida uniformemente entre $[0, 2\pi)$. El vector $\eta(t)$ es ruido y se asume sea un proceso tipo α estable con parámetro de localización cero y características de α desconocidas. Las frecuencias desconocidas están representadas por w_i, y serán estimadas bajo el algoritmo MUSIC.

Para ilustrar el algoritmo se consideraron señales sinusoidales con frecuencias de $\frac{7}{9}\pi$, $\frac{25}{18}\pi$ y $\frac{3}{2}\pi$, contaminadas con ruido α-estable. Para cada realización se tomaron 20 versiones desplazadas en el tiempo del vector muestral $X(t)$ cada una de longitud 200, y se calculó la matriz de correlación para cada uno de los métodos propuestos, dando como resultado en cada uno de los casos una matriz de tamaño 20×20. Se reconstruyó el ejemplo ilustrado por Arce et al [1], bajo el algoritmo MUSIC, donde el parámetro de dispersión del proceso α-estable es 0.2 y el parámetro α=1.2.

A fines de ilustrar el resultado de las componentes de frecuencia, se realizó una normalización de las mismas a 180º, los valores de frecuencias relacionados a esta normalización son w_1=70º, w_2=125º y w_3=135º.

La intención del algoritmo MUSIC es estimar las frecuencias en el subespacio de ruido, a través de la estimación espectral la cual tenderá a infinito para cualquiera

de las p componentes de frecuencias sinusoidales, así los $\hat{p}$ picos más largos de $\hat{P}(w)$ muestran cuales son los valores estimados de las componentes de frecuencia.

En el primer ejemplo, se consideró que el número de componentes de la señal es conocido $p = 3$, es decir, según el algoritmo descrito en la sección 6.2 , el paso 3 es obviado.

La figura 7.2 muestra la señal de prueba sometida a distintos tipos de ruido. En el primer caso observe la figura 7.2(a) la señal en estudio esta contaminada con ruido tipo α-estable con un parámetro de α=2, es decir, un ruido de tipo gaussiano. En el segundo caso la señal de prueba esta contaminada con ruido impulsivo que corresponde a una distribución tipo α-estable con parámetro α=1.2, note en la figura 7.2(b) la presencia de impulsos en la señal.

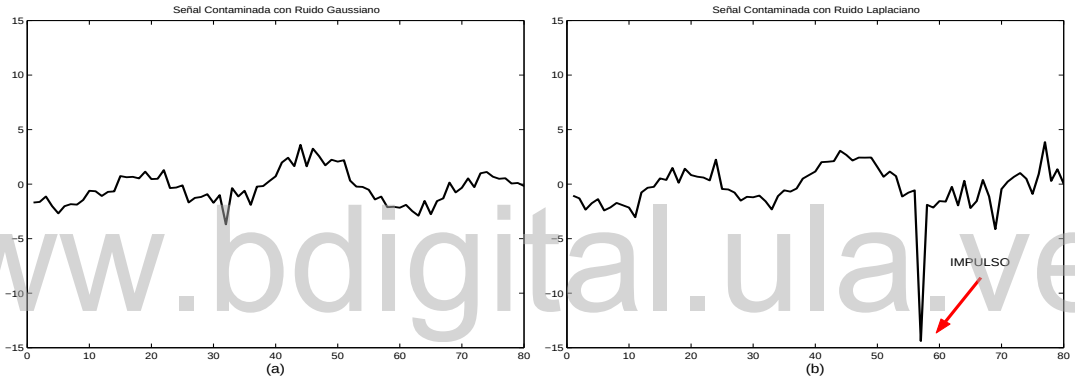

Figura 7.2: (a) Señal contaminada con ruido Gaussiano; (b) Señal contaminada con ruido impulsivo

Para comprobar la robustez de los métodos de correlación descritos, se aplicó el algoritmo de MUSIC tradicional, y luego se sustituyeron los autovectores de la matriz tradicional por los autovectores generados por las matrices robustas.

Las figuras. 7.3 y 7.4 muestran el desempeño de la matriz de correlación bajo el algoritmo MUSIC, donde el número de señales es conocido. Mientras que la señal de prueba esta contaminada con ruido de tipo gaussiano el algoritmo MUSIC utilizando la matriz de correlación tradicional funciona de manera eficiente, pero cuando se encuentra con impulsos, estos puntos dispersos hacen que ocurra un mal calculo de matriz de correlación tradicional y por lo tanto refleja componentes de frecuencia mal ubicados. La figura 7.3(b) muestra el comportamiento de MUSIC usando la matriz de correlación

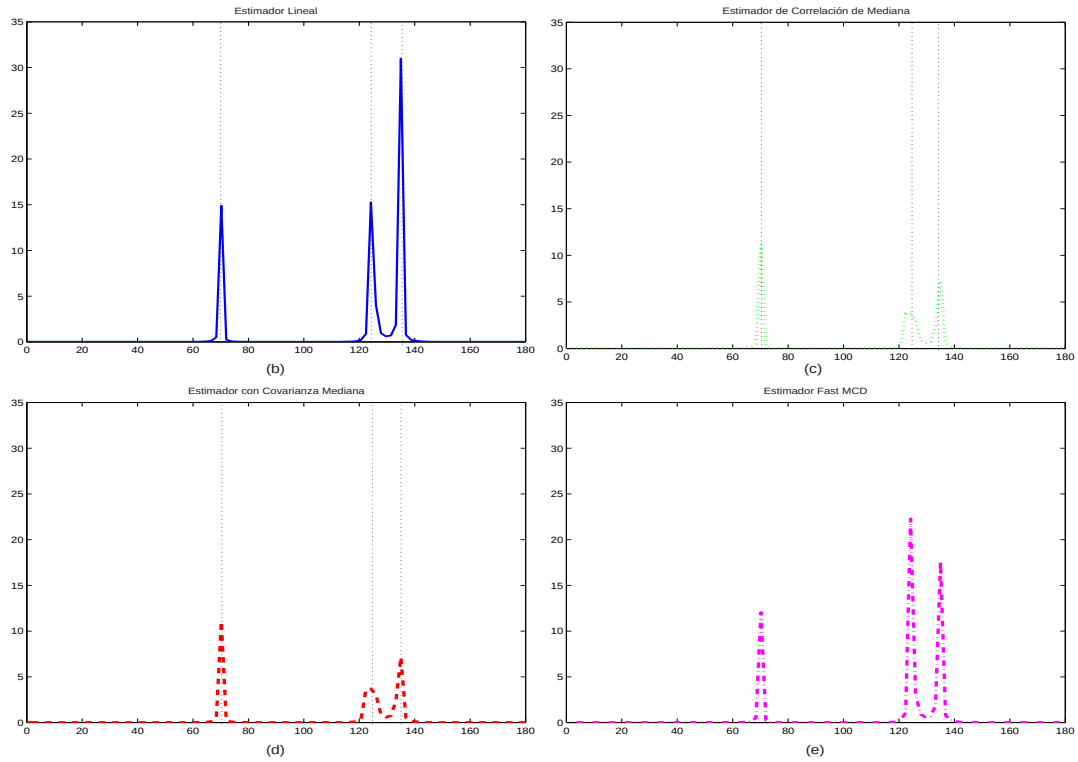

Figura 7.3: Espectro de MUSIC de una señal contaminada con ruido gaussiano usando estimadores b)Lineal, Autocorrelación Mediana Muestral, (d)Autocovarianza Mediana Muestral,(e) MCD

tradicional cuando la señal esta contaminada con ruido de tipo gaussiano, puede notar que estima de manera eficiente las componentes de frecuencias ubicadas en 70°, 125° y 135°. Observe además la Figura 7.4(b), que representa el espectro de MUSIC cuando la señal de prueba esta contaminada con ruido impulsivo, puede notar que en presencia de impulsos el algoritmo de MUSIC usando el estimador Lineal no funciono como el caso Gaussiano, ya que se pierde la resolución entre las componentes de frecuencia que se encuentran ubicadas cerca una de la otra, como es el caso de 125° y 135°, donde los picos que representan el espectro de MUSIC tienden a unirse dando una idea distorsionada de los valores verdaderos de componentes de frecuencia de la señal. Observe ahora la Figura 7.3 (c) y compare con la Figura 7.3 (c), note que ambas figuras muestran el espectro de MUSIC calculado usando la Correlación basada en Mediana Ec.(4.13), la primera cuando la señal esta contaminada con ruido gaussiano, el espectro dibujado muestra la estimación de las componentes de frecuencia de la señal pero de manera muy débil en

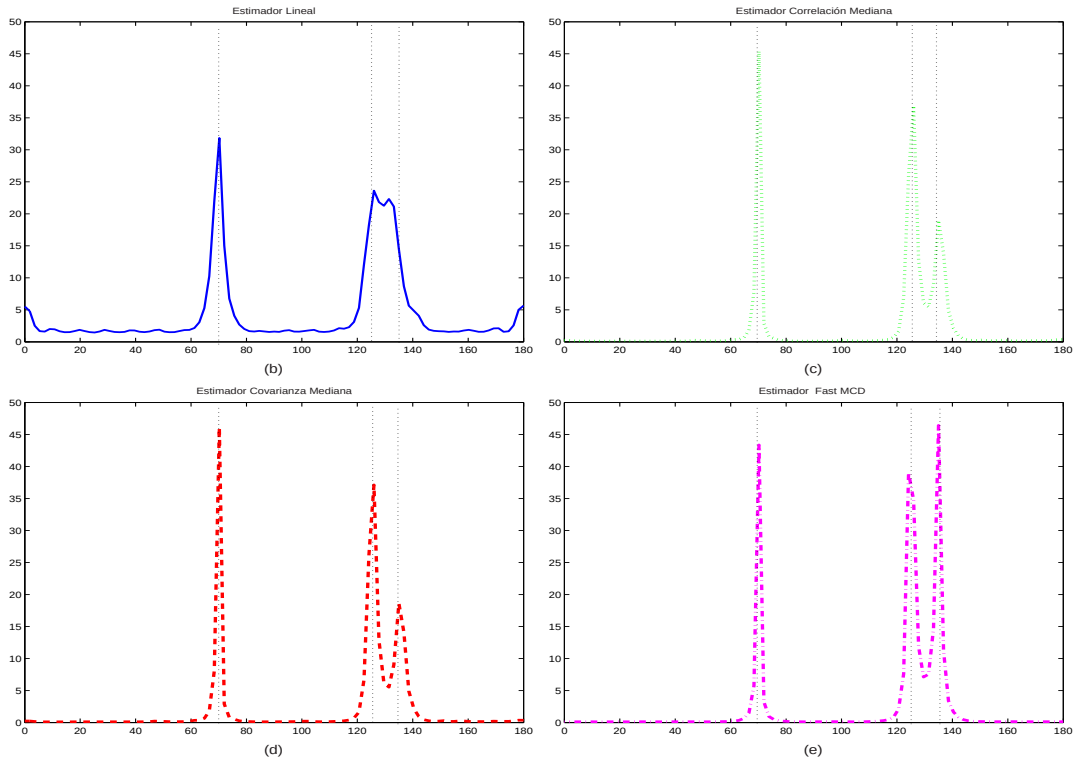

Figura 7.4: Espectro de MUSIC de una señal contaminada con ruido impulsivo tipo α-estable con α=1.2, usando estimadores (b)Lineal, (c)Correlación Mediana, (d) Covarianza Mediana,(e)MCD

comparación de la Figura 7.3 (c) donde la señal esta contaminada con ruido impulsivo, y muestra claramente los picos que indican donde están ubicadas dichas componentes. Las Figura 7.3 (d) y Figura 7.4 (d) representan el algoritmo de MUSIC usando la Covarianza Basada en Mediana Ec.(4.14) y los resultados son equivalentes al caso calculado con la Correlación Basada en Mediana, ambos métodos, basados en el modelo laplaciano, funcionan eficientemente cuando los datos están modelados bajo este criterio, pero dan resultados débiles para casos donde no existe impulsividad en los datos. Cuando MUSIC fue calculado usando la matriz de Correlación de Determinante Mínimo (MCD), tanto para el caso gaussiano como para el laplaciano funciono de manera eficiente, es decir, el grado de impulsividad de los datos no influyen directamente para el calculo de la matriz de correlación bajo este método, lo que hace que MUSIC no se vea afectado en estas condiciones.

En el segundo ejemplo, se consideró que el número de componentes de la señal

es desconocido, según el algoritmo descrito en la Sección 6.2, el número de señales será estimado usando el criterio de la Descripción de la Minima Longitud (MDL).

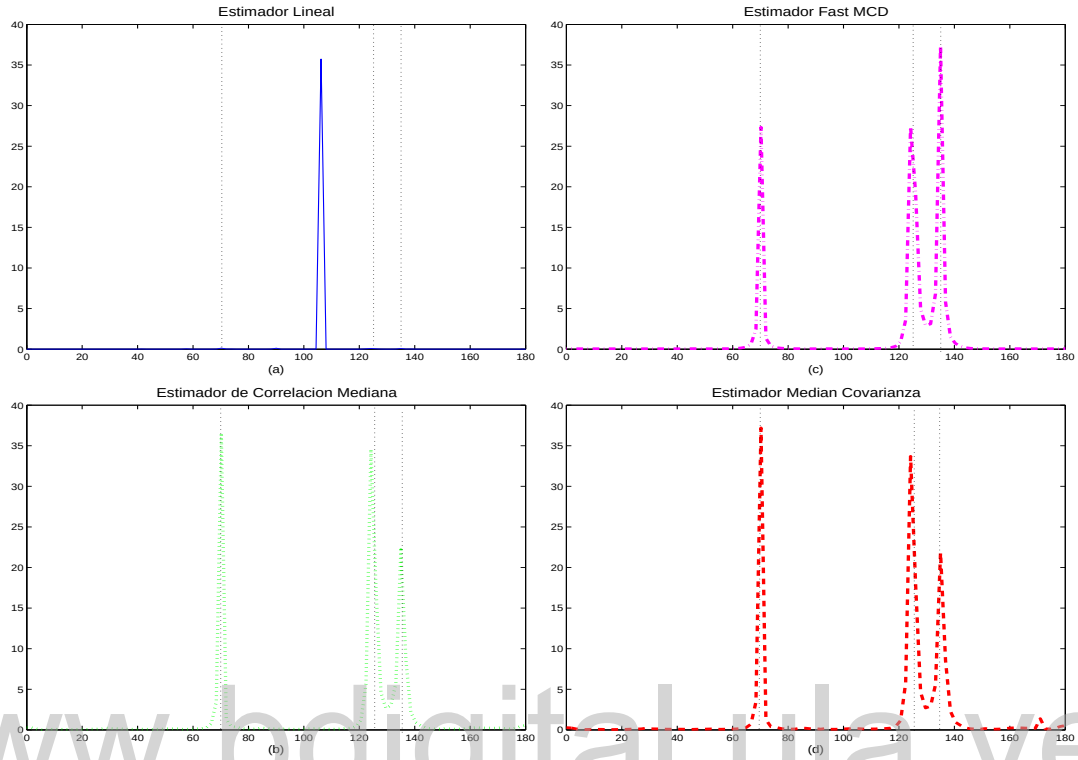

Figura 7.5: Espectro de MUSIC de una señal contaminada con ruido impulsivo tipo α-estable con α=1.2, usando MDL para estimar el número de componentes de frecuencia de la señal. (a)Estimador Lineal, (b)Estimador MCD, (c)Estimador de Correlación Mediana, (d)Estimador de Covarianza Mediana

La Figura 7.5 representa la simulación de los resultados del algoritmo MUSIC usando los métodos de correlación robustos propuestos, donde la señal esta contaminada con una cantidad considerable de impulsividad, en este caso α=1.2. Tome en cuenta que cada uno de los algoritmos usados para esta prueba depende del calculo de la matriz de correlación, un mal calculo puede generar un resultado equivocado del número de señales usando el MDL, por lo tanto, MUSIC generará una estimación equivocada en la ubicación de las componentes de frecuencia de la señal. Como un algoritmo es dependiente del otro, es muy importante la precisión en el calculo de la matriz de correlación.

Se tomó un valor de α lo suficientemente bajo que indicara un grado de impulsividad considerable, tal que se pueda demostrar la robustez de los métodos de correlación

propuestos en [1] y [4]. Por ser una señal aleatoria, se considero la ley de los grandes números, haciendo 500 repeticiones del calculo del espectro de MUSIC, y se tomo el promedio, tal que se pudiera obtener un resultado mas preciso.

Observe las Figura 7.5 (b), Figura 7.5(c), Figura 7.5(d), donde se muestra el espectro de MUSIC, donde en el algoritmo de MUSIC original se sustituyeron los autovectores de la matriz de Correlación Tradicional por los autovectores de las matrices de Correlación Mediana Muestral, Covarianza Mediana Muestral, y de Correlación con Determinante Mínimo (MCD), respectivamente. Alli se refleja la robustez de los métodos descritos, se muestra la estimación eficiente de las componentes de frecuencias que contiene la señal contaminada. Los espectros de MUSIC usando los métodos robustos de correlación muestran claramente los picos ubicados en 70º, 125º y 135º, vea en la Figura 7.5(b) para MUSIC calculado con los autovectores de la Autocorrelación Mediana Muestral; en la Figura 7.5(c) el espectro de MUSIC usando los autovectores de la Autocovarianza Mediana Muestral; y en la Figura7.5(d) MUSIC utilizando los autovectores de la Correlación Basada en MCD. Mientras que la Figura 7.5 (a) muestra el espectro de MUSIC generado por la matriz de Correlación Tradicional, se puede notar claramente que las verdaderas componentes de la señal se pierden, dando como resultado una sola componente ubicada en 110º, esto se debe a la presencia de datos impulsivos que hacen que el cálculo de correlación sea equivocado, y por lo tanto bajo el algoritmo MUSIC se haga una mala estimación de las componentes de frecuencia de la señal. Considere que en este caso el algoritmo de MUSIC depende de la estimación de número de señales, entonces el resultado equivocado del espectro generado por la matriz de Correlación Tradicional puede deberse a un mala del número de señales, para este caso particular si bajo el criterio de MDL se estimó una sola componente sinusoidal en la señal, bajo esta consideración el algoritmo de MUSIC generó la ubicación de esta componente; o el número de componentes de la señal fue bien estimado, pero el espectro de MUSIC falló en la estimación de ubicación de las componentes de la señal por posibles malas estimaciones de matriz de correlación y por lo tanto mal calculo en los autovectores.

7.3. Normalización de Datos de Microarray de ADNc

El algoritmo de normalización descrito en la sección 6.1 fue aplicado a mediciones de expresión genética procedentes de una base de datos disponible en Internet en [23]. Los datos son específicamente el resultado de múltiples repeticiones del experimento del microarray realizados en genes de abejas.

Con el fin de comparar el desempeño de las matrices de correlación robustas se implementó el algoritmo de normalización por correlación a los datos del microarray . Este algoritmo se implementó sustituyendo la correlación tradicional utilizado para el cálculo del parámetro de normalización a dado por la ecuación 6.7, por las correlaciones robustas descritas en los capitulos 2 y 3. Específicamente el valor del parámetro de normalización a usando la Correlación Basada en Mediana es:

$$\tilde{a} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}|Y_i|\right)\cdot MED(|Y_i| \diamond \operatorname{sgn}(Y_i)X_i|_{i=1}^{n})}{\left(\frac{1}{n}\sum_{i=1}^{n}|X_i|\right)\cdot MED(|X_i| \diamond |X_i||_{i=1}^{n})), \tag{7.3}}$$

el cual contiene en el numerador la Correlación Mediana Muestral y en el denominador la Autocorrelación Mediana Muestral.

De igual manera la ecuación que describe al parámetro a, utilizando la Covarianza Basada en Mediana es:

$$\hat{a} = \frac{MED(|Y_i|\,|_{i=1}^{n})\cdot MED(|Y_i| \diamond \operatorname{sgn}(Y_i)X_i|_{i=1}^{n})}{MED(|X_i|\,|_{i=1}^{n})\cdot MED\left(|X_{i,j}| \diamond \operatorname{sgn}(X_{i,j})X_{1,i}|_{i=1}^{n}\right).} \tag{7.4}$$

Observe que en la Ec. (7.4) el parámetro de normalización es proporcional a Covarianza Mediana Muestral e inversamente proporcional a la Autocovarianza Mediana Muestral.

Los resultados de la normalización entre un par de repeticiones de microarray se muestra la Figura 7.6. En tal figura, se muestran cuatro gráficas que despliegan la transformación de los datos usando los métodos de correlación tradicional, correlación mediana muestral, covarianza mediana muestral y la correlación basada en MCD. Adicionalmente, se muestra en cada una de las gráficas la recta $y = x$. Se observa en las cuatro gráficas de la Figura 7.6 que los datos normalizados tienden a la recta $y = x$ en

mayor proporción, comparado con la tendencia de los datos no normalizados. Desde el punto de vista de los experimentos de microarray, este debe ser el comportamiento de los datos normalizados debido a que el proceso de normalización intenta que los valores generados por experimentos idénticos produzcan los mismos resultados [9].

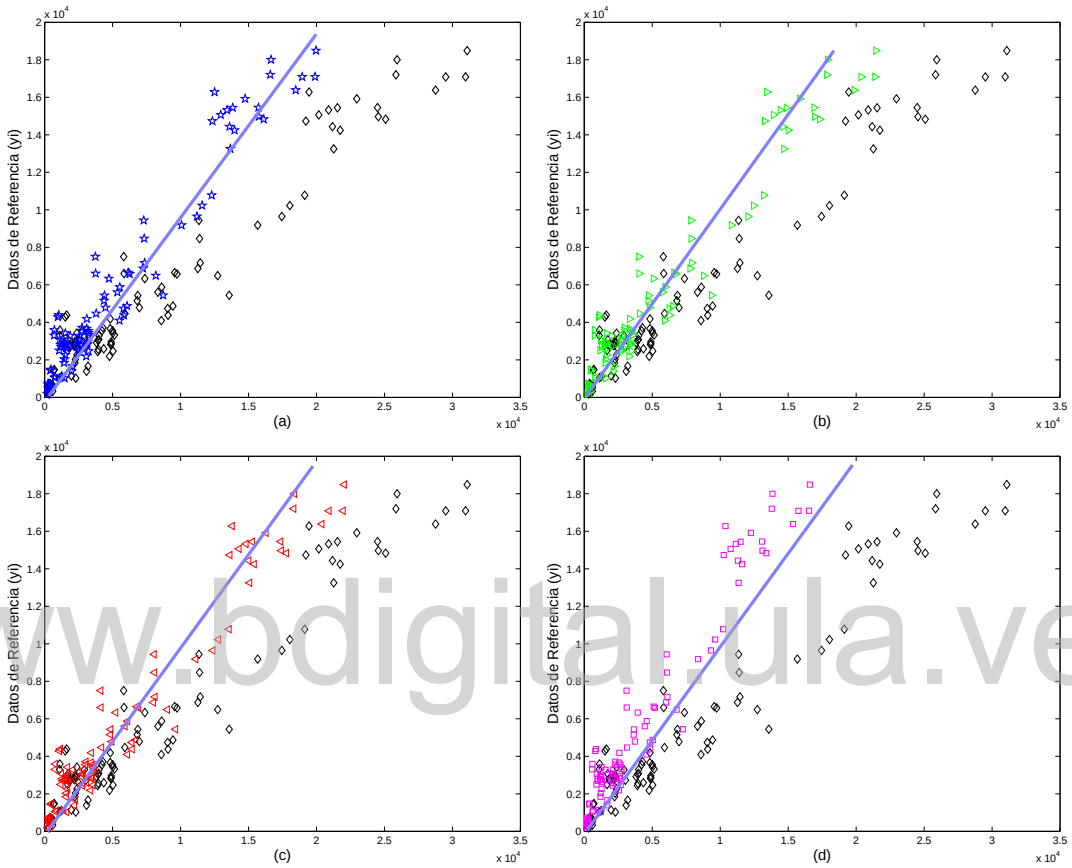

Figura 7.6: Comparación de los datos originales ♢ con los datos normalizados utilizando la matriz de (a) Correlación Tradicional ⋆, (b) Correlación Basada en Mediana ▷, (c) Covarianza Basada en Mediana ◁, (d) Correlación Basada en MCD

El criterio usado para cuantificar el desempeño de los métodos de normalización fue el error absoluto medio (MAE). En la Figura 7.7 se muestra el MAE de los métodos de normalización seleccionados para esta aplicación, comparados con el MAE obtenido de los datos originales, es decir,no normalizados (linea punteada). Se observa en todos los métodos usados, que el MAE de los datos después de la normalización, es en general siempre menor que el MAE generado por los datos no normalizados. Los métodos que tienen un desempeño mejor a lo largo de todas las repeticiones son el de correlación

mediana muestral y el de covarianza mediana muestral, debido a que el error producido por estos métodos no supera el error generado por los datos originales.

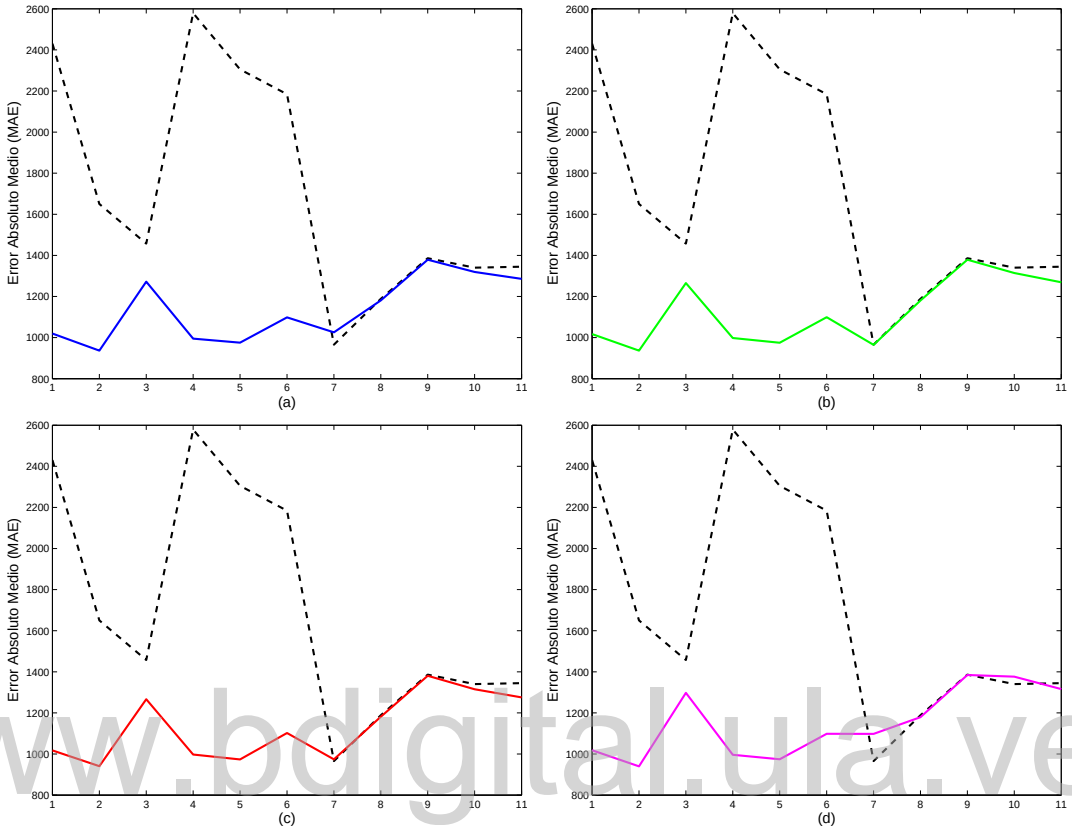

Figura 7.7: Gráfica del Error Absoluto Medio de los datos de referencia con respecto a cada observación de los datos originales (linea punteada), y a los datos normalizados (linea continua) usando (a) Correlación Tradicional; (b) Correlación Mediana Muestral; (c) Covarianza Mediana Muestral; (d) Correlación Basada en Mediana

Finalmente en la Figura 7.8 se muestra cuatro gráficos que representan las nubes de puntos producidas por los datos originales y los datos normalizados usando los métodos anteriormente descritos. En estas gráficas están la tendencia de los datos procedentes de la repetición que genera un error mayor en comparación a los datos no normalizados, es decir, la repetición 7 del experimento. Se observa en dicha gráfica que los métodos de correlación mediana muestral y covarianza mediana muestral generan datos normalizados que se adaptan mejor a la recta $y = x$ en comparación con la correlación tradicional y la correlación basada en MCD. Esto indica que los métodos de correlación mediana muestral y covarianza mediana muestral presentan mayor robustez

ante tendencias no lineales de los datos tal como muestra las nubes de puntos de los datos no normalizados desplegadas también en la Figura 7.8.

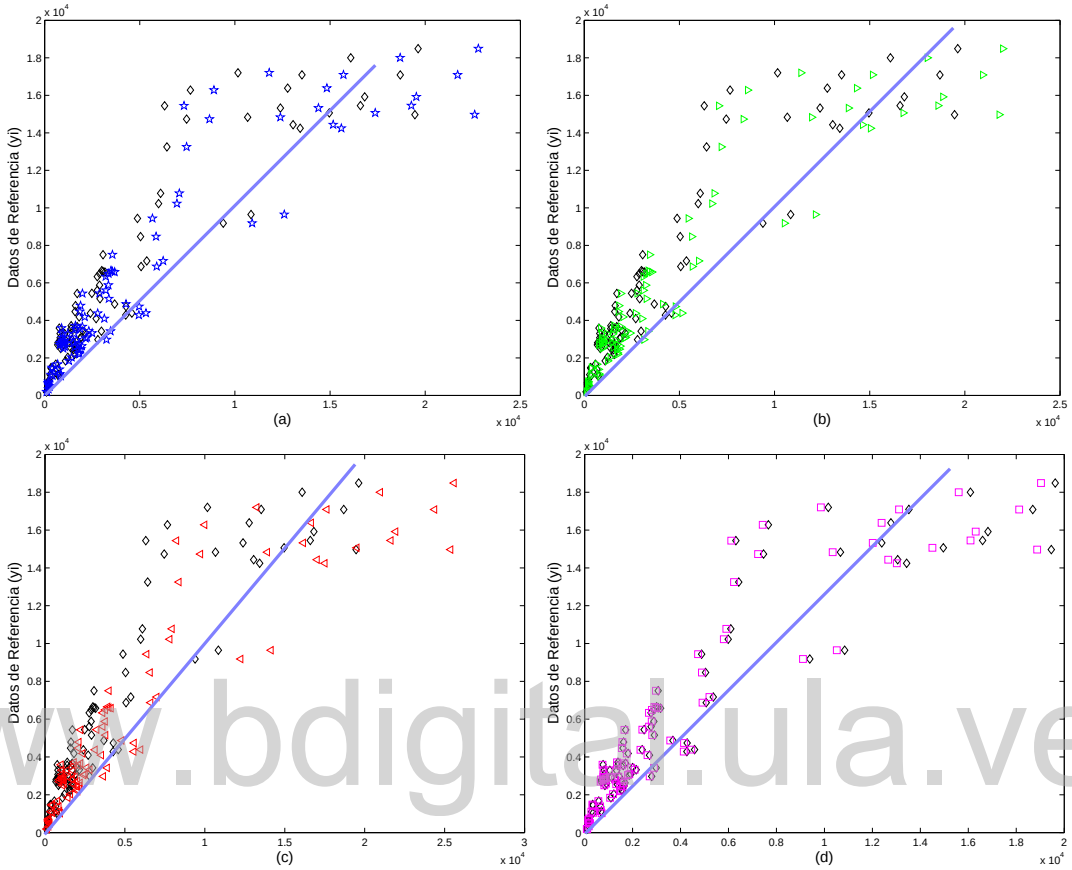

Figura 7.8: ÇASO IMPULSOÇomparación de los datos originales ♢ con los datos normalizados utilizando la matriz de (a) Correlación Tradicional ⋆, (b) Correlación Basada en Mediana ▷, (c) Covarianza Basada en Mediana ◁, (d) Correlación Basada en MCD

Capítulo 8

Conclusiones y Recomendaciones

- En la práctica el ruido que contamina la señal es de tipo impulsivo y no de tipo gaussiano como se ha considerado hasta ahora en el procesamiento de señales. Cuando una señal contiene datos dispersos, no es suficiente obtener la matriz tradicional modelada bajo la distribución gaussiana, es necesario considerar técnicas robustas que funcionen eficientemente ante ruido impulsivo. Bajo esta consideración, se demostró que en presencia del ruido impulsivo los métodos robustos propuestos por Arce y Rousseauw proporcionan una estimación precisa de correlación. Esto origina nuevas expectivas en la investigación ampliando el campo del uso de técnicas no lineales, mediante la utilización de estimación de parámetros estadísticos para el desarrollo de métodos de procesamiento de señales.

- Existen innumerables aplicaciones en la literatura que dependen notablemente del cálculo de correlación muestral. Para que dichas aplicaciones funcionen eficientemente es indispensable hacer una estimación de la correlación, considerando que si la correlación esta mal estimada, es muy probable que el algoritmo que se este ejecutando produzca resultados pobres. En este trabajo la robustez de los métodos de correlación estudiados se mostró a través de simulaciones en las aplicaciones de carácter práctico Descripción de la Mínima Longitud (MDL), Clasificación de Múltiples Señales (MUSIC), y Normalización de Datos de Microarray de ADNc, los cuales dependen del cálculo de correlación muestral.

- Los distintos métodos de estimación de correlación se calcularon siguiendo distintos procedimientos, según el caso. Para la estimación de la matriz de Correlación Tradicional el procedimiento y los cálculos se consideran simples, en general se reduce a la sumatoria del producto de las muestras. El cálculo de las Matrices de Correlación Basadas en Mediana, tienen una complejidad intermedia; y presenta nuevos conceptos de Mediana Ponderada para cualquier valor real. El cálculo de Correlación Basada en MCD, es significativamente más complejo que el anterior, y se basa en el cálculo iterativo de la distancia de Mahalanobis, obteniendo el subconjunto para el cual el determinante es mínimo.

- Se mostró la robustez del método propuesto por Arce [1] en distintas aplicaciones y para distintos niveles de ruido. Específicamente, se mostró que la Correlación y Covarianza Basadas en Mediana son más eficientes en comparación con la Correlación Tradicional, cuando el ruido que contamina la señal es impulsivo. Sin embargo, los métodos Robustos de Correlación Basados en la Mediana, no son muy eficientes cuando el ruido que contamina la señal es de naturaleza gaussiana, a diferencia de la Correlación Tradicional que no presenta problemas en este caso.

- La robustez del método propuesto por Rousseauw [4] fue mostrada en distintas aplicaciones y considerando señales contaminadas con ruido de diferentes grados de impulsividad. Los resultados mostraron que la Correlación Basada en MCD es más eficiente en comparación con la Correlación Tradicional, cuando el ruido que contamina la señal es impulsivo. Sin embargo, a diferencia de los métodos de Correlación Basados en Mediana, la Correlación Basada en MCD funciona de manera eficiente cuando el ruido que contamina la señal es de tipo gaussiano.

- La Correlación y Covarianza Basadas en Mediana, al igual que la Correlación Basada en MCD, calculan de manera eficiente en presencia de ruido impulsivo el número de componentes de frecuencia de una señal usando el algoritmo MDL.

- Los resultados empíricos muestran que en presencia de ruido impulsivo tanto la Correlación Mediana Muestral y la Covarianza Mediana Muestral, como la Correlación obtenida mediante el MCD son mucho más eficientes que la correlación tradicional en cuanto a la estimación de las componentes espectrales usando

el algoritmo MUSIC.

- El método de Correlación Basada en MCD aunque funciona de manera eficiente para los tipos de contaminación de ruido estudiados, a nivel computacional es poco eficiente, en comparación con los otros métodos.

- Se considera que existe la necesidad de modelar métodos de correlación que sean lo suficientemente robustos tal que funcionen de manera eficiente ante cualquier intensidad de ruido.

- Los métodos de correlación robustos descritos presentan limitaciones. Los métodos de Correlación Basados en Mediana, reducen su efectividad a un rango de impulsividad en el ruido. El método Basado en MCD presenta limitaciones a nivel computacional. Entonces se considera necesario modelar nuevos métodos de correlación que tomen en cuenta la eficiencia ante distintos tipos de ruido, y que sean eficiente considerando la velocidad en el cálculo.

- Finalmente como continuación de este trabajo se recomienda ampliar el estudio comparativo de los distintos métodos de correlación usando otros tipos de aplicaciones.

Bibliografía

[1] Arce, G. Li, Y. Median power and median correlation theory. *IEEE Transactions on Signal Processing.* vol.50, pp.2768-2776,.

[2] Briceño, J. *Principios de las Comunicaciones.* Consejo de Publicaciones, Universidad de Los Andes, Mérida, 1era edición.

[3] Gonzalez, J. *"Robust techniques for wireless communications in non-gaussian environments," Ph.D. dissertation.*

[4] Rousseeuw, P. y Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics, Vol. 41, pp 212-223.*

[5] Kailath, T., Wax, F. Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustic, Speech, and Signal Processing.* vol.ASSP-33, pp.387-392.

[6] Hayes, M. *Statistical Digital Signal Processing and Modeling.* Jhon Wiley & Sons, Inc,.

[7] J.H Zar. *Biostadistical Analisys.* 4ta edición Upper Saddle River,NJ: Prentice-Hall, 1999.

[8] J. Rissanen. Modeling by shortest data description. *Automatica, Vol. 14, pp 465-471.*

[9] Lee, R Gu, Z Clarke, R Wang, Y, Lu, J. Iterative normalization of cdna microarray data. *IEEE Transactions on Inormation Tecnology in Biomedicine.* Vol. 6 Nº1, Marzo 2002.

[10] Rodriguez, J. Moreno, A. Mariño, J., Vallverdú, F. *Tratamiento Digital de la señal, Una introducción Experimental*. Edicions UPC, España.

[11] Ziemer, R. Tranter, W. *Principles of Communicatons*. Houghton Nifflin Company, Boston.

[12] Scheaffer, R. Mendenhall, W., Wackerly, D. *Estadistica Matematica con Aplicaciones*. Grupo Editorial Iberomaericana, 2da edición.

[13] Panagiotis, G.Kyriakakis y Panayiotis, G. Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise. *IEEE Transactions on Multimedia, Vol. 1, Nº 3, September*, 1999.

[14] Preben, K. Alpha-stable distributions in signal processing of audio signals. 1998.

[15] Proakis, J. Manolakis, D. *Tratamiento Digital de Señales*. Prentice-Hall, Madrid, 3era edición.

[16] Papoulis, A. *The Fourier integral and its applications*. Mc Graw Hill, 1978.

[17] Wozencraft, J. Jacobs, I. *Principles of Communications Engineering*. Jhon Wiley & Sons, Inc, 1965.

[18] Oppenheim, A. *Señales y sistemas*. Prentice-Hall, México.

[19] Arce, G. A general weighted median filter structure admitting negative weigths. *IEEE Transactions on Signal Processing*. vol.46, pp.3195-3205.

[20] Mcd estimator and robust distances (fast-mcd). `www.agoras.ua.ac.be/`.

[21] Mahalanobis-distance. `www.answers.com/topic`, Enero, 2005.

[22] Shao, M. and Nikias, C. *Signal Processing with Alpha stable Distributions and Aplications*. Wily, NY, 1993.

[23] European informatics institut. `www.ebi.ac.uk/arrayexpress/`. Marzo, 2005.

[24] Haykin, S. *Adaptative Filter Theory*. Prentice Hall information and System Sciences Series, NJ, 3era edicion, 1996.