Revista de la Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín
V 5 N 2 Julio-Diciembre de 2016 • ISSN-e 2357-5749 • Artículo Investigación • Páginas 12 a 37
DOI: https://doi.org/10.15446/rev.fac.cienc.v5n2.50666

ROBUSTIFICACIÓN DE LA CARTA DE CONTROL MULTIVARIADA $\sqrt{|\mathbf{S}|}$ EN LA FASE I DE CONTROL[a]

ROBUSTIFICATION OF MULTIVARIATE CONTROL CHART $\sqrt{|\mathbf{S}|}$ IN CONTROL PHASE I

EDWIN DUGARTE PEÑA[b],

NELFI GERTRUDIS GONZÁLEZ ÁLVAREZ[c]

Recibido 15-05-2015, aceptado 02-09-2016, versión final 04-09-2016.
Artículo Investigación

RESUMEN: En este artículo se estudia la robustificación de la carta basada en la raiz cuadrada de la varianza muestral generalizada $\sqrt{|\mathbf{S}|}$ para el control de la variabilidad de un proceso normal bivariado, en la etapa 1 de la Fase I de control, construida con observaciones sobre subgrupos racionales y utilizando los estimadores robustos MVE, MCD, estimador S. Estas cartas se comparan con la carta usual basada en el estimador insesgado muestral $\mathbf{S}$ de la matriz de covarianza $\mathbf{\Sigma}_0$, en presencia de outliers provenientes de esquemas de perturbación del tipo contaminación con inflación de $\mathbf{\Sigma}_0$ y contaminación perturbando sólo la correlación. Como medida de desempeño se usa el error cuadrático medio en la estimación de $\mathbf{\Sigma}_0$ y el sesgo absoluto en la estimación de $\sqrt{|\mathbf{\Sigma}_0|}$, sobre los estimadores insesgados para cada uno de estos parámetros, respectivamente, construidos con los subgrupos racionales que quedan después del proceso de depuración realizado en la Fase I y que se consideran como el conjunto de datos que representa el estado de variación estable del proceso.

PALABRAS CLAVE: Control estadístico multivariado, estimación robusta, fase I de control, matriz de covarianzas

ABSTRACT: This article, chart robustification based on the square root of the sample generalized variance $\sqrt{|\mathbf{S}|}$, is studied to control the variability of a bivariate normal process in stage 1 of Phase I control, built with observations for rational subgroups, and using the robust estimators, MVE, MCD and estimator S. These charts, are being compared with the usual chart, based on unbiased sample estimator, $\mathbf{S}$, of the covariance matrix, $\mathbf{\Sigma}_0$, in the presence of outliers from perturbation schemes such as, $\mathbf{\Sigma}_0$ inflation contamination, and contamination affecting only the correlation. We use as performance measure, the mean squared error in estimating $\mathbf{\Sigma}_0$, and absolute bias in the estimate, used $\sqrt{|\mathbf{\Sigma}_0|}$, about the unbiased estimators for each one of

[a]Dugarte Peña, E. & González Álvarez N. G. (2016). Robustificación de la carta de control multivariada $\sqrt{|\mathbf{S}|}$ en la fase I de control *Revista de la Facultad de Ciencias*, 5 (**2**), 12–37. DOI: https://doi.org/10.15446/rev.fac.cienc.v5n2.50666

[b]Profesor Asociado, Escuela de Ingeniería, Universidad Pontificia Bolivariana, Seccional Bucaramanga, edugartep@unal.edu.co.

[c]Profesor Asociado, Escuela de Estadística, Universidad Nacional de Colombia, sede Medellín, ngonzale@unal.edu.co.

the parameters respectively built with the rational subgroups that remain after the purification process undertaken in Phase 1 and are consider as the data group that represent the stage of the stable variation process.

KEYWORDS: Multivariate statistical control, robust estimation, phase I control, covariance matrix

1. INTRODUCCIÓN

Ocurre que en muchos procesos de control estadístico se supone que la dispersión del proceso es constante y es igual a la matriz de covarianzas, $\mathbf{\Sigma}$. Este supuesto no es cierto en general y debe ser validado en la práctica. La variabilidad del proceso se resume en la matriz de covarianzas de dimensión $p \times p$, que contiene $\frac{p \times (p+1)}{2}$ parámetros. La variabilidad total de un conjunto de datos multivariados, se puede medir por una o más de las siguientes opciones: *i)* La Varianza Generalizada, que corresponde con el determinante de la matriz de covarianzas, $|\mathbf{\Sigma}|$. Esta es proporcional a la superficie o el volumen generado por un conjunto de datos. *ii)* La raíz cuadrada de la varianza generalizada y *iii)* La traza de la matriz de covarianzas, $tr(\mathbf{\Sigma})$, que representa la varianza total o la suma de las varianzas de las variables.

El objetivo primordial del control estadístico de procesos es comprender y controlar la variación de una o varias características de calidad. Para el caso multivariado se han formulado varios esquemas para el monitoreo de la matriz de varianzas y covarianzas (Alt, 1985; Alt & Smith, 1988) y para las cuales a su vez también han sido propuestas algunas modificaciones que tratan de mejorar el desempeño de estos procedimientos en la Fase II de control (Alt, 1985; Alt & Smith, 1988; Djauhari *et al.*, 2008).

La implementación de un proceso de control multivariado requiere dos Fases: Una Fase I, que básicamente es una fase de estimación de los parámetros que describen el proceso en su estado de variación estable, y una Fase II de monitoreo en la que se chequea la continuidad del estado de control a lo largo del tiempo, bajo la operación del proceso.

En la Fase I de control, el principal objetivo es obtener una estimación de los parámetros del proceso (univariado o multivariado) que corresponda a su estado de variación estable. En particular, este trabajo se enfoca en la matriz de covarianzas $\mathbf{\Sigma}_0$. La estimación está basada en un conjunto de datos históricos que se supone son representativos del estado de variación estable, es decir, sin la presencia de fuentes o causas asignables de variación, supuesto que puede no ser cierto en general y debe ser validado en la práctica.

Un tipo de datos anómalos, asociados a estados de fuera de control, son los conocidos como outliers u observaciones outliers. Dados los efectos adversos que tienen los subgrupos de outliers en la es-

timación de los parámetros, es necesario hacer uso de métodos que no sean fácilmente afectados por este tipo de observaciones. Estos métodos son denominados robustos, cuyos resultados siguen siendo confiables aún si cierta cantidad de datos están contaminados (Rousseeuw & Leroy, 1987).

Este trabajo pretende evaluar en un proceso normal bivariado, la carta de control para el monitoreo estadístico de la raíz cuadrada de la varianza generalizada, $\sqrt{|\mathbf{\Sigma}_0|}$ con observaciones agrupadas, y el desempeño de tales cartas en la depuración del conjunto de datos de Fase I, construidas usando los estimadores robustos MVE, MCD y S y el estimador insesgado muestral usual, en presencia de outliers provenientes de esquemas de perturbación de $\mathbf{\Sigma}_0$, del tipo contaminación con inflación de esta matriz y contaminación perturbando sólo la correlación. Como medida de desempeño se usa el error cuadrático medio en la estimación de $\mathbf{\Sigma}_0$ y el sesgo absoluto en la estimación de $\sqrt{|\mathbf{\Sigma}_0|}$, sobre los estimadores insesgados para cada uno de estos parámetros, respectivamente, construidos con los subgrupos racionales que quedan después del proceso de depuración realizado en la Fase I y que se consideran como el conjunto de datos que representa el estado de variación estable del proceso, bajo normalidad bivariada, con observaciones de subgrupos racionales, tomando como referencia la carta $\sqrt{|\mathbf{S}|}$ (Alt & Smith, 1988).

2. MARCO TEÓRICO

2.1. Control estadístico multivariado

La variación en un proceso se ha caracterizado a partir de dos tipos de causas: comunes y asignables o especiales. En el primer tipo, se asume que la variabilidad observada es debida a la naturaleza propia del proceso y por tal razón, no puede ser alterada sin que el proceso en sí mismo cambie. En el segundo tipo, se consideran las situaciones inusuales o también las causas de lo que puede y debe ser eliminado del proceso.

Los datos que se utilizan en el monitoreo y en el análisis de control estadístico de procesos pueden ser el resultado de observaciones sobre unidades individuales o de observaciones sobre sugbrupos racionales. Según Davis & Yen (1998), un subgrupo racional debe ser una muestra tomada de forma tal que se maximice la probabilidad de capturar la variabilidad debida a causas comunes y que cualquier variabilidad debida a causas especiales ocurra entre subgrupos.

Tal y como lo presenta Jensen *et al.* (2006), la Fase I del esquema de monitoreo consiste en determinar si los datos históricos indican o no un proceso estable; por lo tanto, se vuelve muy importante descubrir si hay puntos de datos inusuales tales como outliers antes de calcular los límites de control.

Un supuesto que no siempre se verifica apropiadamente es que los datos de la Fase I provienen de un proceso en control y en particular, la presencia de observaciones outliers dentro del conjunto de da-

tos históricos, recolectados en la Fase I, puede conducir a límites de control "inflados" y estimadores sesgados de los parámetros, que luego conducirán a una menor capacidad para detectar cambios en el proceso durante la Fase II. Según Barnett & Lewis (1998) un outlier es un conjunto de datos en una observación (o subconjunto de observaciones) que parece ser inconsistente (desviarse notablemente) con el resto del conjunto de datos. A partir de uno de los principios generales propuestos por Barnett (1979) para la detección de outliers, es posible inferir un concepto de outlier multivariado: "La observación más extrema es aquella $\mathbf{X}_i \in \mathbb{R}^p$ cuya omisión en la muestra $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ produce el mayor aumento incremental en la verosimilitud maximizada bajo el modelo básico que se haya especificado para los datos restantes. Si este aumento es sorprendentemente grande, se declara que $\mathbf{X}_i$ es un outlier". Para el caso de subgrupos racionales y teniendo en cuenta la definición propuesta por Nelson (1988), se puede decir que un outlier es cualquier muestra o subgrupo racional cuyas unidades muestrales se ajustan total o parcialmente a una distribución diferente de la distribución subyacente.

De acuerdo con Jensen *et al.* (2006), los métodos robustos de estimación tienen una clara ventaja sobre los métodos clásicos en que no se ven indebidamente influenciados por los outliers. En consecuencia, son mucho más eficaces en la detección de outliers y en asegurar que los límites de control sean razonables.

2.2. Observaciones contaminantes

De acuerdo con Barnett & Lewis (1998), las observaciones contaminantes pueden ser interpretadas como observaciones que provienen de una distribución diferente de la distribución base a la que se ajustan las observaciones. Señalan que la naturaleza y origen de los outliers puede estar determinada por una forma aleatoria o no explicable y/o por una forma determinística y/o por alguna influencia externa que no guarda relación en lo específico con el proceso que se observa.

Dado lo anterior, si se asume que para una población multivariada más de la mitad de los datos son buenos, es decir, se ajustan a una distribución subyacente F_θ, y por lo tanto una fracción menor del 50 % de los datos no lo son (datos incorrectos o outliers), es decir, no se ajustan al patrón de la mayoría de los datos y provienen de una población diferente y si ε representa la fracción de observaciones que se consideran como outliers generados por una distribución diferente H, entonces según Barnett & Lewis (1998), se define la familia de contaminación a partir de

$$\mathcal{F}_\varepsilon = \{(1-\varepsilon)F_\theta + \varepsilon H : \theta \in \Theta\} \tag{1}$$

donde ε representa la fracción de contaminación.

Para simular la presencia de observaciones contaminantes respecto de la distribución subyacente, se utiliza el concepto de perturbación. Una perturbación se puede interpretar como la suplantación de

información de la distribución subyacente por información de otra u otras distribuciones diferentes. Las perturbaciones han sido clasificadas como perturbaciones difusas y perturbaciones localizadas. Según Tatum (1997), una perturbación se clasifica como difusa si se verifica que las unidades contaminantes u outliers se extienden por todas las muestras o subgrupos racionales con la misma probabilidad de ocurrencia. Señala que la perturbación localizada es aquella en la que se produce un impacto en todas las unidades de una o varias muestras o subgrupos racionales.

2.3. Estimación robusta

Según Huber *et al.* (2008), la estimación robusta consiste en encontrar un ajuste "robusto", que sea similar a la forma que se hubiera encontrado sin los outliers.

Para construir estimadores robustos de $\mathbf{\Sigma}_0$ combinando la información de m muestras independientes, se requiere sustituir los estimadores de máxima verosimilitud de la matriz de covarianzas por sus análogos robustos. En Todorov & Filzmore (2009) y Djauhari *et al.* (2008) se presentan algunas formas para resolver esta necesidad. En este artículo se optó por utilizar el promedio de las matrices sobre la muestra final de la Fase I.

De acuerdo con Jensen *et al.* (2006) los métodos de estimación robustos pueden ser utilizados con base en dos enfoques diferentes. El primer enfoque es el uso de los estimadores robustos en lugar de los estimadores clásicos. El segundo enfoque consiste en utilizar los estimadores robustos para identificar y eliminar los outliers y luego usar los estimadores clásicos de los restantes datos "buenos". Aunque los estimadores robustos tienen menor eficiencia ya que sólo utilizan un poco más de la mitad de los puntos disponibles, esta no es tan crucial, ya que los estimadores robustos eventualmente serán reemplazados por estimadores clásicos.

Para esta investigación, los métodos de estimación robustos se utilizan bajo el segundo enfoque. La idea entonces es identificar y eliminar los subgrupos con datos anómalos a partir de la utilización de un estimador robusto y luego usar los estimadores clásicos de los restantes datos "buenos".

2.4. Estimadores robustos multivariados

De acuerdo con Barnett & Lewis (1998), la esencia del concepto de robustez, en cuanto a la presencia de outliers y los métodos robustos, se centra en la necesidad que se tiene de proceder con seguridad a pesar de la presencia de ellos.

El tema de estimación robusta y/o detección de outliers ha sido investigado por muchos autores, entre otros Rousseeuw (1985), quien introdujo bajo normalidad multivariada el estimador del elipsoide de mínimo volumen (MVE) y el estimador de matriz de covarianzas de determinante mínimo

(MCD) propuesto incialmente por Rousseeuw (1984) y posteriormente mejorado por Rousseeuw & Van Driessen (1999) para detección de outliers.

Según Jensen *et al.* (2006), en aplicaciones de control de calidad en Fase I, el MCD y el MVE se utilizan directamente para determinar outliers multivariados y por lo tanto se vuelve más importante que sean lo suficientemente precisos.

En la revisión realizada se encuentra que Chenouri *et al.* (2009) usaron versiones reponderadas del estimador MCD (RMCD) para hacer seguimiento a observaciones individuales multivariadas en la Fase II de control; Chenouri & Variyath (2010) realizaron estudios para evaluar el desempeño de las cartas T2 de Hotelling's para observaciones individuales con estimadores reponderados MCD, MVE y estimadores S en la Fase II de control y de igual manera, Variyath & Vattathoor (2013) usaron versiones reponderadas de los estimadores MCD y MVE (RMCD y RMVE) para monitorear observaciones multivariadas en la Fase I de control. Aunque probaron que las versiones reponderadas son más eficientes que las versiones MVE y MCD, estas pruebas se realizaron bajo el contexto de observaciones individuales.

A continuación se hace una presentación más detallada de cada uno de los estimadores robustos, incluyendo los estimadores S.

2.4.1. Estimadores S

Los estimadores S de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ fueron introducidos por Davies (1987) y además estudiados por Lopuhaä (1989), Rousseeuw & Yohai (1984), citados por Rousseeuw & Leroy (1987) introdujeron en el campo de la regresión la clase de estimadores S. Lopuhaä & Rousseeuw (1991) presentan una generalización de estos estimadores para localización y covarianza multivariada como sigue:

Definición Sea $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ con $n \geq p+1$, un conjunto de n observaciones en $\mathbb{R}^p$ y ρ: $\mathbb{R} \to [0, \infty)$ que satisface las condiciones

- $\rho(.)$ es simétrica, su derivada $\psi(.)$ es continua y $\rho(0) = 0$
- Existe una constante finita $c_0 > 0$ tal que $\rho(.)$ es estrictamente creciente en $[0, c_0]$ y constante en $[c_0, \infty)$.

Un estimador S multivariado de localización y covarianza es la solución $\boldsymbol{\theta}_n = (\mathbf{t}_n, \mathbf{C}_n)$ al problema de minimizar el determinante $|\mathbf{C}|$ con $\mathbf{t}_n \in \mathbb{R}^p$ y $\mathbf{C}_n$ en el conjunto de todas las matrices simétricas definidas positiva de tamaño p; es decir, dada una función ρ, el problema se centra en determinar:

$$(\mathbf{t}_n, \mathbf{C}_n) = \arg\min_{\mathbf{t},\mathbf{C}} \det(\mathbf{C}), \quad \text{tal que} \quad \frac{1}{n}\sum_{i=1}^{n} \rho\left(\sqrt{(\mathbf{x}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{t})}\right) = b_0 \tag{2}$$

La constante b_0 que puede asumir valores $0 < b_0 < \sup \rho$ puede ser elegida de acuerdo con una distribución supuesta. La relación anterior puede ser escrita también como

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(d_i/c\right) = b_0, \quad \text{donde} \quad d_i = \sqrt{(\mathbf{x}_i - \mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{t})} \tag{3}$$

Al estimador $\mathbf{t}$ se le denomina un M-estimador de localización, mientras que $\mathbf{C}$ es un M-estimador de dispersión multivariado. Si estos parámetros se estiman simultáneamente, ellos son llamados S-estimadores. Desde este último punto de vista, sólo se considera S-estimadores ya que son altamente resistentes a los outliers para una función ρ elegida apropiadamente. La calidad de los S-estimadores depende de la función ρ. Se recomienda utilizar la función ρ biponderada de Tukey, ya que ésta no varía con el cambio de dimensión.

Para el cálculo del estimador robusto S, se decidió entre los estimadores robustos S trabajar con el estimador robusto basado en el algoritmo S-FAST propuesto por Salibian-Barrera & Yohai (2006). Para su cálculo, Todorov & Filzmore (2009) y Todorov (2012) proporcionan en el paquete R, en la librería **rrcov** la función **CovSest()** que calcula estimaciones-S multivariadas de localización y escala usando ponderadores de la función bisquare de Tukey's y un algoritmo fast.

2.4.2. Estimadores MVE (elipsoide de mínimo volumen)

La estimación busca encontrar el elipsoide de volumen mínimo que cubre un subconjunto de al menos h puntos de datos. Los subconjuntos de tamaño h son llamados conjuntos-medios porque h se elige a menudo para ser justo mayor o igual que la mitad de los n puntos de datos.

El estimador MVE, de localización multivariado $\mathbf{t}$ corresponde al centro del elipsoide de volumen mínimo que cubre al menos el 50 % de los puntos; el estimador de covarianza $\mathbf{C}$ corresponde al volumen de dicho elipsoide multiplicado por un factor de corrección para obtener consistencia (Rousseeuw & Leroy, 1987) ; luego el estimador MVE de localización y dispersión no corresponde con el vector de medias muestrales y la matriz de covarianzas muestral de un conjunto-medio particular.

Rousseeuw & van Zomeren (1990) presentan un estimador que tiene la propiedad de ser afín equivariante. Sea $\mathbf{X}_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ un conjunto de n datos con $\mathbf{x}_i \in \mathbb{R}^p$, seleccionadas desde una distribución normal p-variada. El estimador MVE de localización y de la matriz de covarianzas, está definido como el par $(\mathbf{t}, \mathbf{C})$, donde $\mathbf{t}$ es un vector de dimensión p y $\mathbf{C}$ es una matriz simétrica definida positiva tal que el determinante de $\mathbf{C}$ es minimizado sujeto a

$$\#\{i; (\mathbf{x}_i - \mathbf{t})^T\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{t}) \leq a^2\} \geq \left[\frac{n + p + 1}{2}\right] = h \tag{4}$$

Donde el símbolo # corresponde con el número de puntos que satisfacen la condición y a^2 es una constante que puede tomarse igual a $\chi^2_{(0{,}5,p)}$ cuando se espera que la mayoría de los datos provengan

de una distribución normal.

El estimador MVE es afín equivariante debido a que la imagen de un elipsoide a través de una transformación afín no singular; es decir, de la forma $\mathbf{x} \rightarrow \mathbf{A}\mathbf{x} + \mathbf{b}$, es también un elipsoide con volumen igual a $|\mathbf{A}| \times$ vol, donde vol es el volumen original y dado que $|\mathbf{A}|$ es una constante, el tamaño relativo del elipsoide no cambia bajo transformaciones afín (Rousseeuw & Leroy, 1987).

2.4.3. Estimador de la matriz de covarianzas de mínimo determinante (MCD)

Un procedimiento alternativo de estimación al MVE con alto punto de ruptura es el estimador basado en la covarianza de mínimo determinante (MCD), propuesto por primera vez por Rousseeuw (1984) y posteriormente mejorado por Rousseeuw & Van Driessen (1999).

Los estimadores MCD son intuitivamente atractivos ya que un valor pequeño del determinante corresponde con dependencias lineales de los datos en el espacio p-dimensional cercanas. Esto se debe a que un determinante pequeño corresponde con un valor propio pequeño que sugiere una dependencia casi lineal que a su vez sugiere que hay un grupo de puntos que son similares entre sí (Jensen *et al.*, 2006).

Suponga que se tiene una muestra $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)'$ de n observaciones desde una población p-variada con media $\boldsymbol{\mu}$ y matriz de dispersión $\boldsymbol{\Sigma}$. El estimador MCD se determina a partir de las h observaciones cuya matriz de covarianzas tenga el menor determinante, con $n/2 \leq h < n$. En cada caso, el estimador de localización es el promedio de estas h observaciones. Si $y_{MCD} = \{i_1, \ldots, i_h\}$ denota los índices de las h observaciones, entonces

$$\hat{\boldsymbol{\mu}}_{MCD} = \frac{1}{h} \sum_{i \epsilon y_{MCD}} \mathbf{Y}_i \tag{5}$$

así entonces, el estimador resultante de localización es el vector de medias muestral de los puntos que se encuentran en el conjunto-medio (halfset).

El estimador de la matriz de covarianzas es la correspondiente matriz de covarianzas. La matriz de covarianzas obtenida en cada caso se multiplica por un factor de consistencia y otro factor de corrección de sesgo para muestras finitas, para que el estimador sea consistente con el modelo normal y sea insesgado para muestras pequeñas. Adicionalmente, se refina seleccionando aquellos puntos cuya distancia de Mahalanobis a la media inicial, usando la matriz de covarianzas obtenida, no sea demasiado grande, y recalculando la media y la matriz de covarianzas; se tiene que:

$$\hat{\boldsymbol{\Sigma}}_{MCD} = \frac{c(h)s(h,n,p)}{h-1} \sum_{i \epsilon y_{(MCD)}} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_{MCD})(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_{MCD})^T \tag{6}$$

donde $c(h)$ es una constante de proporcionalidad que hace a $\hat{\boldsymbol{\Sigma}}_{MCD}$ consistente cuando la distribución de y sea elíptica simétrica y unimodal; es decir, consistente con el modelo normal multivariado, ver **?**, referenciados en Todorov & Filzmore (2009).

Si $y \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ entonces

$$c(h) = \frac{h/n}{P(\chi^2_{p+2} < \chi^2_{p,1-h/n})}$$

donde $\chi^2_{(p,\alpha)}$ denota el $\alpha\,\%$ punto de corte de la distribución χ^2_p.

La segunda constante de proporcionalidad $s(h,n,p)$ sirve como factor de corrección de sesgo para muestras finitas de $\hat{\boldsymbol{\Sigma}}_{MCD}$. El valor real de este factor depende también de n y p; fue obtenido por Pison *et al.* (2002) a través de una combinación de la simulación Monte Carlo y la interpolación paramétrica, bajo el supuesto de que $s(h,n,p) \to 1$ cuando $n \to \infty$ para p fijo. Cerioli *et al.* (2009).

Las ecuaciones (5) y (6) definen las estimaciones robustas MCD de localización y dispersión.

Para el cálculo de los estimadores MVE y MCD Todorov & Filzmore (2009) y Todorov (2012) proporcionan el paquete en R de métodos robustos en la librería **rrcov** en el que se hace una descripción de estimación robusta de localización y escala y análisis robusto multivariado con alto punto de ruptura usando los estimadores MVE y "FAST MCD" respectivamente. En este paquete se presenta la función **CovMve()** que calcula el estimador robusto multivariado de localización y escala utilizando para ello el algoritmo Fast MVE Todorov & Filzmore (2009) y la función **CovMcd()** que calcula el estimador robusto multivariado de localización y escala basado en el algoritmo Fast MCD (Rousseeuw & Van Driessen, 1999).

2.5. Precisión de los estimadores

La precisión de un estimador expresa la cercanía de las estimaciones respecto del parámetro de interés. Ante un escenario en el que se tenga que comparar la precisión de varios estimadores clásicos o robustos, se puede hacer uso del error cuadrático medio (ECM)y el sesgo absoluto. Para el caso multivariado, Köksoy (2006) citado por Gomes *et al.* (2012) propuso la aglutinación de las funciones de error cuadrático medio de cada componente del vector de parámetros de interés, que podrían ser o no ponderadas. A partir de este resultado y de acuerdo con Gomes *et al.* (2012) si se supone que $\boldsymbol{\theta}$ es un vector p-dimensional y $\hat{\boldsymbol{\theta}}$ es un estimador p variado de $\boldsymbol{\theta}$, entonces

$$ECM_T\left[\hat{\boldsymbol{\theta}}\right] = \sum_{j=1}^{p}\left[E\left(\hat{\theta}_j - \theta_j\right)^2\right] = E\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right] \tag{7}$$

De otra parte, una propiedad interesante del ECM es que se puede descomponer como la suma de

la varianza del estimador más su sesgo al cuadrado, es decir

$$ECM_{\hat{\theta},\theta_0} = Var(\hat{\theta}) + \text{Sesgo}^2(\hat{\theta})$$

ahora, si se llama $B = \text{Sesgo}(\hat{\theta})$ entonces

$$B = E(\hat{\theta}) - \theta_0$$

y el sesgo absoluto es $|B|$, así entonces, por la ley de los grandes números se tendrá que:

$$\text{cuando} \quad m \to \infty, \widehat{E(\hat{\theta}_N)} \to E(\hat{\theta}) = \mu_{\hat{\theta}}$$

Por lo anterior, para N grande,

$$\widehat{E(\hat{\theta}_N)} - \theta_0 \simeq B$$

y por lo tanto

$$|\widehat{E(\hat{\theta}_N)} - \theta_0| \simeq |B|.$$

Para el caso tratado en este artículo, N representa el número de veces que se simula la Fase I, siendo $\theta = \sqrt{|\Sigma_0|}$.

Tal y como se presenta en Tatum (1997), Schoonhoven *et al.* (2011) y Schoonhoven & Does (2012), en esta investigación se tomó como medida de desempeño de la carta de control robustificada en Fase I el error cuadrático medio, así como también el sesgo absoluto. Lo anterior se sustenta en que el fin último del proceso abordado es tener una estimación de los parámetros del proceso que sea representativa del estado estable. Otras medidas como la proporción de outliers detectados o la probabilidad de falsa alarma son más relevantes cuando se trabaja sobre cartas de control para datos individuales, para esto ver por ejemplo Vargas (2003), Chenouri *et al.* (2009), Yañez *et al.* (2010), Variyath & Vattathoor (2013), Jones-Farmer *et al.* (2014).

3. CARTAS DE CONTROL

Se requiere determinar la estimación de $\mathbf{\Sigma}_0$ y los límites de control de la carta $\sqrt{|\mathbf{\Sigma}_0|}$ en Fase I, tanto en su forma no robusta como en su forma robusta. Dado que al reemplazar los estimadores insesgados por estimadores robustos no se conoce la distribución en muestras finitas, es necesario usar simulación para el cálculo de los límites de control.

3.1. Carta $\sqrt{|\mathbf{S}|}$

La carta $\sqrt{|\mathbf{S}|}$ se basa en la raíz cuadrada de la varianza generalizada de la muestra, donde $\mathbf{S}$ es la matriz de covarianzas $p \times p$ de la muestra, definida como $\mathbf{S} = \dfrac{1}{n-1} \sum_{i=1}^{n} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^T$. La

carta resultante puede ser considerada como un análogo multivariado de la carta univariada de la desviación estándar de la muestra, carta S.

En la Fase I se utilizan sólo los dos primeros momentos de $\sqrt{|\mathbf{S}|}$ y la propiedad de que la mayor parte de la distribución de probabilidad de $\sqrt{|\mathbf{S}|}$ se encuentra en el intervalo (por aproximación del Teorema de Límite Central),

$$E(|\mathbf{S}|^{1/2}) \pm 3\sqrt{V(|\mathbf{S}|^{1/2})} \tag{8}$$

A partir de lo que presenta Anderson (1984), se tiene que:

$$|\mathbf{S}|^r \sim |\mathbf{\Sigma}_0|^r (n-1)^{-pr} \left[\prod_{k=1}^{p} \chi^2_{n-k}\right]^r \tag{9}$$

así entonces,

$$E(|\mathbf{S}|^r) = |\mathbf{\Sigma}_0|^r (n-1)^{-pr} E\left\{\left[\prod_{k=1}^{p} \chi^2_{n-k}\right]^r\right\} \tag{10}$$

que es equivalente a,

$$E(|\mathbf{S}|^r) = |\mathbf{\Sigma}_0|^r (n-1)^{-pr} 2^{pr} \prod_{k=1}^{p} \left[\frac{\Gamma(\frac{n-k}{2}+r)}{\Gamma(\frac{n-k}{2})}\right] \tag{11}$$

Si se llama $b_r = (n-1)^{-pr} 2^{pr} \prod_{k=1}^{p} \left[\dfrac{\Gamma(\frac{n-k}{2}+r)}{\Gamma(\frac{n-k}{2})}\right]$, entonces

$$E(|\mathbf{S}|^r) = |\mathbf{\Sigma}_0|^r b_r \tag{12}$$

de donde, un estimador insesgado de $|\mathbf{\Sigma}_0|^r$ es

$$\widehat{|\mathbf{\Sigma}_0|^r} = \frac{|\mathbf{S}|^r}{b_r} \tag{13}$$

así entonces, con $r = 1/2$, de (7) y llamando $b_3 = b_{1/2}$, se tiene que

$$E\left(|\mathbf{S}|^{1/2}\right) = |\mathbf{\Sigma}_0|^{1/2} b_3 \tag{14}$$

Ahora, sabiendo que

$$V(|\mathbf{S}|^{1/2}) \quad = \quad E(|\mathbf{S}|) - [E(|\mathbf{S}|^{1/2})]^2$$

se llega al siguiente resultado:

$$V\left(|\mathbf{S}|^{1/2}\right) = |\mathbf{\Sigma}_0|(b_1 - b_3^2) \tag{15}$$

por lo tanto, si se reemplaza (9) y (10) en (3) se obtiene

$$|\boldsymbol{\Sigma}_0|^{1/2} b_3 \pm 3\sqrt{|\boldsymbol{\Sigma}_0|(b_1 - b_3^2)}$$

es decir

$$|\boldsymbol{\Sigma}_0|^{1/2}\left(b_3 \pm 3\sqrt{b_1 - b_3^2}\right) \tag{16}$$

Para el cálculo de los límites de control de esta carta en Fase I, se requieren estimadores insesgados de $|\boldsymbol{\Sigma}_0|^{1/2}$. Suponiendo que se tienen m muestras independientes de tamaño n procedentes de un proceso normal bivariado, sea $\mathbf{S}_k$ la matriz de covarianzas muestral $p \times p$ de la muestra k, para $k = 1, 2, \ldots, m$, donde $\mathbf{S}_k = \dfrac{1}{n-1}\sum_{j=1}^{n}\left(\mathbf{X}_{kj} - \bar{\mathbf{X}}_k\right)\left(\mathbf{X}_{kj} - \bar{\mathbf{X}}_k\right)^T$.

Un estimador insesgado de $|\boldsymbol{\Sigma}_0|^{1/2}$ es $\dfrac{\dfrac{1}{m}\sum_{k=1}^{m}|\mathbf{S}_k|^{1/2}}{b_3}$, por lo tanto si se llama

$$|\mathbf{S}_0|^{1/2} = \frac{1}{m}\sum_{k=1}^{m}|\mathbf{S}_k|^{1/2}, \tag{17}$$

se tiene que

$$\widehat{|\boldsymbol{\Sigma}_0|^{1/2}} = \frac{|\mathbf{S}_0|^{1/2}}{b_3} \tag{18}$$

es un estimador insesgado de $|\boldsymbol{\Sigma}_0|^{1/2}$. Luego, los límites estimados para la carta $\sqrt{|\mathbf{S}|^{1/2}}$ están dados por

$$\frac{|\mathbf{S}_0|^{1/2}}{b_3}\left(b_3 \pm 3\sqrt{b_1 - b_3^2}\right) \tag{19}$$

de acuerdo con lo anterior, los límites de control 3σ $(\alpha/2 = 0{,}00135)$ para la carta $|\mathbf{S}|^{1/2}$ en Fase I se pueden determinar mediante:

$$UCL = \frac{|\mathbf{S}_0|^{1/2}}{b_3}\left(b_3 + 3\sqrt{b_1 - b_3^2}\right) \tag{20}$$

$$CL = |\mathbf{S}_0|^{1/2} \tag{21}$$

$$LCL = \text{máx}\left\{0, \frac{|\mathbf{S}_0|^{1/2}}{b_3}\left(b_3 - 3\sqrt{b_1 - b_3^2}\right)\right\} \tag{22}$$

Es importante tener en cuenta que si $\mathbf{S}$ es definida positiva, entonces $|\mathbf{S}|^{1/2} > 0$ y como no es significativo tener un límite de control inferior que sea negativo, en el caso que esto ocurra se ajustará el LCL a cero.

3.2. Versión robusta de la carta $\sqrt{|\mathbf{S}|}$

Sea $\mathbf{S}_R$ el estimador robusto de $\boldsymbol{\Sigma}_0$. Siguiendo a (12), se tiene que $\mathrm{E}(|\mathbf{S}_R|^{1/2}) = |\boldsymbol{\Sigma}_0|^{1/2} b_{3R}$ y $\mathrm{E}(|\mathbf{S}_R|) = |\boldsymbol{\Sigma}_0| b_{1R}$, donde b_{1R} y b_{3R} corresponden a la versión robusta de las constantes b_1 y b_3.

Una vez más, basados en m muestras independientes de tamaño n de un proceso normal bivariado, sea $\mathbf{S}_{R,k}$ el estimador robusto de $\boldsymbol{\Sigma}_0$ en la $k-$ésima muestra, $k = 1, \ldots, m$. Entonces, un estimador insesgado de $|\boldsymbol{\Sigma}_0|^{1/2}$ es

$$\widehat{|\boldsymbol{\Sigma}_0|^{1/2}} = \frac{|\mathbf{S}_{0,R}|^{1/2}}{b_{3R}} \tag{23}$$

donde

$$|\mathbf{S}_{0,R}|^{1/2} = \frac{1}{m} \sum_{k=1}^{m} |\mathbf{S}_{R,k}|^{1/2}, \tag{24}$$

y similarmente, un estimador insesgado para $|\boldsymbol{\Sigma}_0|$ es

$$\widehat{|\boldsymbol{\Sigma}_0|} = \frac{|\mathbf{S}_{0,R}|}{b_{1R}} \tag{25}$$

con

$$|\mathbf{S}_{0,R}| = \frac{1}{m} \sum_{k=1}^{m} |\mathbf{S}_{R,k}|, \tag{26}$$

A partir de (19) los límites robustos para la carta $\sqrt{|\mathbf{S}|}$ se determinan mediante

$$\frac{|\mathbf{S}_{0,R}|^{1/2}}{b_{3R}} \left(b_{3R} \pm 3\sqrt{b_{1R} - b_{3R}^2} \right), \tag{27}$$

donde b_{1R} y b_{3R} se determinan siguiendo a (12), es decir,

$$\frac{|\mathbf{S}_{0,R}|}{|\Sigma_0|} = \frac{\frac{1}{m}\sum_{k=1}^{m} |\mathbf{S}_{R,k}|}{|\Sigma_0|} \cong b_{1R}, \qquad \frac{|\mathbf{S}_{0,R}|^{1/2}}{|\Sigma_0|^{1/2}} = \frac{\frac{1}{m}\sum_{k=1}^{m} |\mathbf{S}_{R,k}|^{1/2}}{|\Sigma_0|^{1/2}} \cong b_{3R}, \quad \text{cuando} \quad m \to \infty, \tag{28}$$

Los valores de b_{1R} y b_{3R} para los diferentes estimadores tenidos en cuenta se pueden observar en la Tabla 1.

4. ESTUDIO DE SIMULACIÓN

El objetivo central de este trabajo es proponer un procedimiento Fase I para el control multivariado de la matriz de covarianzas que sea robusto a outliers y que proporcione un subconjunto de datos históricos con los que se pueda estimar con el menor sesgo y la mejor precisión posible en los parámetros que definen la medida de dispersión multivariada $\boldsymbol{\Sigma}$, y $\sqrt{|\boldsymbol{\Sigma}|}$, considerando la carta de control de la raíz cuadrada de la varianza muestral generalizada.

Se asumió que $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$ son el vector de medias y matriz de covarianzas bajo la distribución subyacente[d]. y $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$ son las correspondientes a la distribución contaminante, donde, para el caso que nos interesa, $\boldsymbol{\Sigma}_1$ estará definida como $\Sigma_1 = \begin{pmatrix} \delta_x\sigma^2_{0,x} & 0 \\ 0 & \delta_y\sigma^2_{0,y} \end{pmatrix}$ siendo δ_x y δ_y el factor de inflación en las respectivas componentes de la matriz de covarianzas.

Las estimaciones de la matriz de covarianzas y los correspondientes límites de control se determinaron con base en 100000 simulaciones. El software utilizado fue el paquete R Core Team (2013).

Los factores de simulación que se sometieron a variación fueron los siguientes:

1. Número de subgrupos (m): Este factor tiene como objetivo establecer el efecto del número de subgrupos en el proceso de estimación. Se tomaron valores de m =20, 30 y 40.

2. Tamaño de los subgrupos (n): Este factor tiene como objetivo establecer el efecto del tamaño del subgrupo en el proceso de estimación. Se tomaron valores de n =10, 15 y 20.

3. Entornos de contaminación. Se manejaron dos entornos de contaminación: contaminación por perturbación localizada y contaminación por perturbación difusa.

4. Esquemas de contaminación. Se manejaron tres esquemas de contaminación:

   - La varianza de una de las características de calidad en k de los m subgrupos racionales crece de $\sigma^2_{0,x}$ a $\delta_x\sigma^2_{0,x}$ (o $\sigma^2_{0,y}$ a $\delta_y\sigma^2_{0,y}$) para δ_x, δ_y >1 si el otro permanece fijo.
   - La varianza de las dos características de calidad en k de los m subgrupos racionales crece de $\sigma^2_{0,x}$ a $\delta_x\sigma^2_{0,x}$ y de $\sigma^2_{0,y}$ a $\delta_y\sigma^2_{0,y}$ para δ_x, δ_y =1.5 y 3.
   - Afectando la correlación entre las características de calidad.

De acuerdo con Maronna *et al.* (2006), con respecto a la caracterización real de procesos correlacionados, la decisión de considerar un proceso bivariado no correlacionado se sustenta en el hecho que los estadísticos usados para las estimaciones de la dispersión (MCD, MVE y S) en la carta considerada, son afín equivariantes, lo que significa que se comportan adecuadamente bajo transformaciones afines de los datos, es decir, cambiar la escala de medición o localización no debe afectar las propiedades del estimador.

Para el caso del estimador usual bajo los diferentes formatos de contaminación se utilizaron los valores exactos bajo normalidad para las constantes b_1 y b_3. Tal y como lo presenta Tatum

[d] A partir del supuesto $\boldsymbol{\Sigma}_0 = I_2$ se tendrá que las variables en el proceso son independientes; no están correlacionadas. una propuesta que también se podría tener en cuenta es la presentada en Vargas & Lagos (2007) en la que toman $\boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. En esta propuesta Vargas and Lagos asumen ρ=0.5

(1997), para proporcionar estimadores insesgados de la $|S|$ para el caso iid Normal, los valores proporcionados por cada uno de los tres estimadores robustos deben ser divididos por una constante de normalización, en este caso, las constantes b_{1R} y b_{3R}. Tales constantes se obtuvieron a partir de un millón de simulaciones. (Ver Tabla 1)

Tabla 1: Constantes b_1 y b_3 para la carta $\sqrt{|\mathbf{S}|}$

	Usual	MCD	MVE	S-FAST
		n = 10		
b_1	0.8894720	1.1236391	0.7015406	0.5538867
b_3	0.8891952	0.9466571	0.7618410	0.6751277
		n = 15		
b_1	0.9293023	1.0450593	0.8195529	0.7150831
b_3	0.9288599	0.9650706	0.8564492	0.7998158
		n = 20		
b_1	0.947328	1.0388239	0.8714618	0.7910446
b_3	0.947327	0.9779878	0.8953765	0.8554401

4.1. Resultados y análisis de resultados

Precisión y sesgo de los estimadores Usual, MCD, MVE y S-FAST, con la carta de control $\sqrt{|\mathbf{S}|}$ sobre la muestra final de la Fase I, asumiendo contaminación localizada. (ver Figuras 1 y 2):

- Bajo condiciones de factor de inflación δ_x y/o δ_y=1.5, cuando el número de subgrupos contaminantes es bajo, $k \leq 2$, tamaño de muestra pequeño, n =10, y el número de subgrupos m =20, el procedimiento Fase I diferencia claramente al estimador MCD con alguna ventaja frente a los demás estimadores robustos y el estimador usual, al compararse en cuanto al $ECM(\widehat{\mathbf{\Sigma}})$, ventaja que desaparece o disminuye al incrementar el tamaño de muestra, el número de subgrupos y/o el número de subgrupos contaminados. Llama la atención que con este estimador se presenta la peor condición en cuanto al sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$ en todos los escenarios observados.

- Bajo condiciones de factor de inflación δ_x y/o δ_y=1.5, con respecto al estimador S-FAST se encontró que su desempeño en cuanto a $ECM(\widehat{\mathbf{\Sigma}})$ y sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$ mejora a medida que n >10 y m >20, sin embargo, al ser comparado con los demás estimadores, la ventaja que inicialmente alcanza desaparece en la medida que n y m se hacen cada vez más grandes y/o crece el número de subgrupos contaminados.

- Para el estimador MVE su $ECM(\widehat{\mathbf{\Sigma}})$ y sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$ no presentan algún mejor desempeño dado que en la generalidad de los resultados, estos se muestran entre los de los demás estimadores.

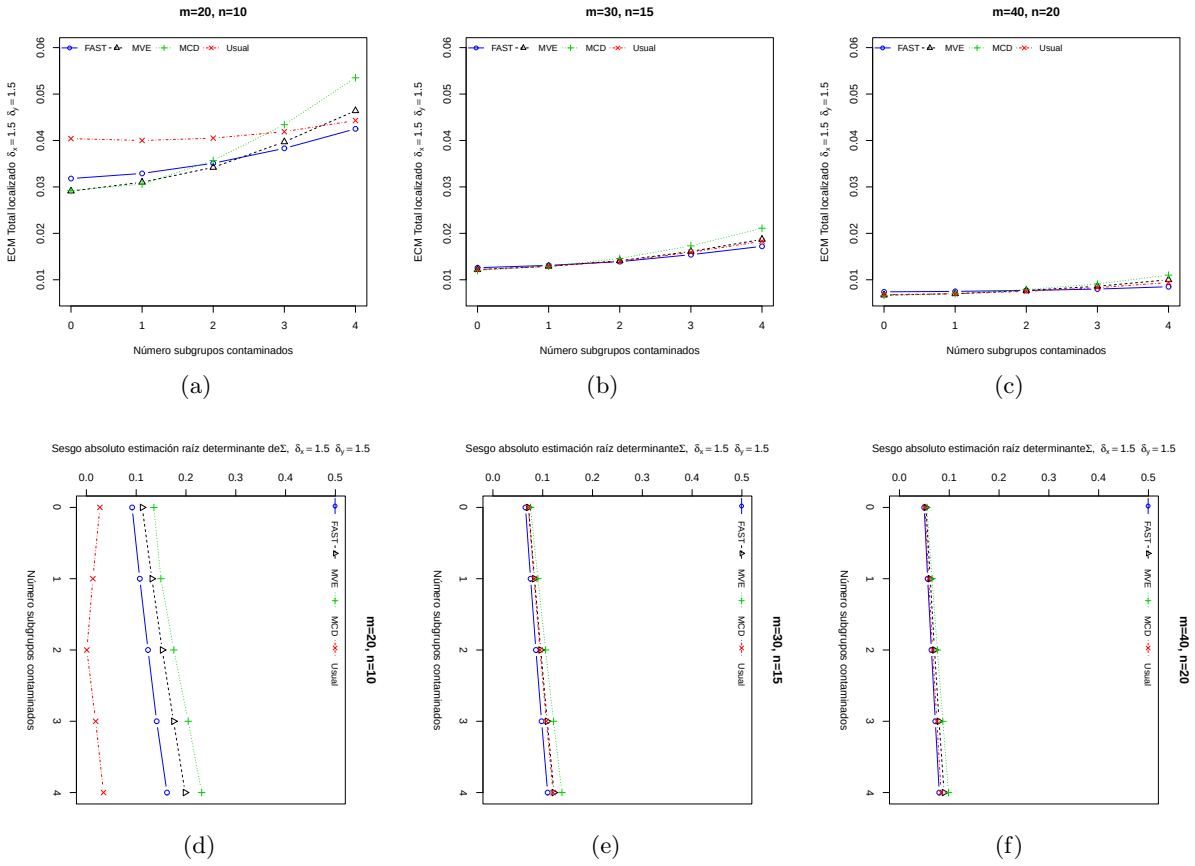

Figura 1: Resultados bajo contaminación localizada con δ_x = δ_y=1.5

- Cuando el factor de inflación de la varianza se elevó a δ_x y/o δ_y=3, bajo contaminación localizada, en todos los escenarios se presentó peor el estimador MCD tanto en $ECM(\widehat{\mathbf{\Sigma}})$ como en el sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$.

- Bajo condiciones de factor de inflación δ_x y/o δ_y=3, el procedimiento Fase I usando el estimador S-FAST muestra alguna ventaja bien diferenciada frente a los demás estimadores tanto en $ECM(\widehat{\mathbf{\Sigma}})$ como en el sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$, cuando n >10 y m >20, sin embargo, la ventaja que inicialmente alcanza desaparece en la medida que n y m se hacen cada vez más grandes y/o crece el número de subgrupos contaminados.

- Cabe resaltar que en la medida que crece el número de subgrupos, el tamaño de la muestra y el número de subgrupos contaminados, bajo contaminación localizada afectando sólo parámetros de varianza, el $ECM(\widehat{\mathbf{\Sigma}})$ de los diferentes estimadores robustos tiende a ser muy similar entre sí y con el estimador Usual, razón por la que se puede asumir que bajo estos términos no se presentan resultados positivos para los casos robustos comparados con el procedimiento

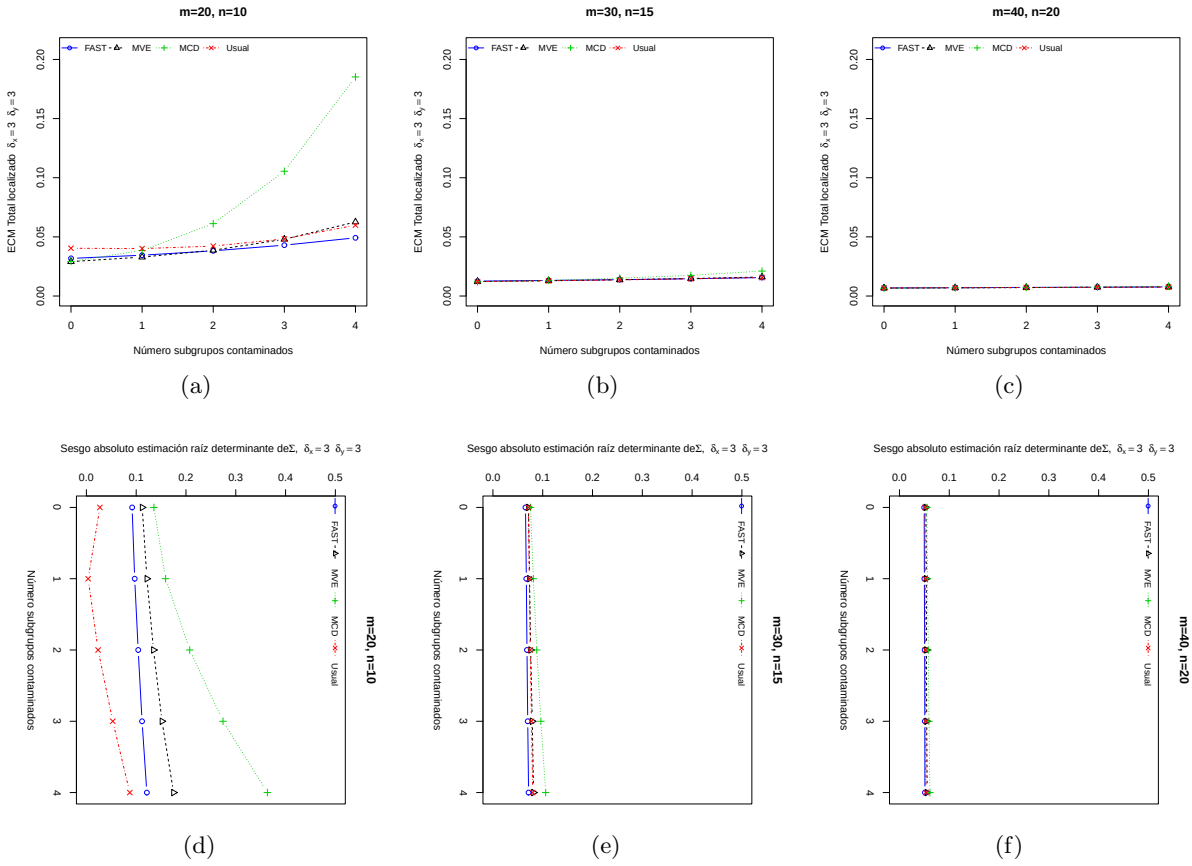

Figura 2: Resultados bajo contaminación localizada con δ_x =3, δ_y =3

basado en el estimador Usual.

Al evaluar el $ECM(\widehat{\boldsymbol{\Sigma}})$ y el sesgo absoluto de la estimación de $\sqrt{|\boldsymbol{\Sigma}|}$ con los estimadores Usual, MCD, MVE y S-FAST, con la carta de control $\sqrt{|\mathbf{S}|}$ sobre la muestra final de la Fase I, y asumiendo contaminación difusa, se hace evidente a partir de las Figuras 3 y 4, cada una de las siguientes observaciones:

- Se observó que el procedimiento Fase I usando el estimador Usual, para muestras pequeñas, n=10, en todos los escenarios, presenta la peor condición en cuanto al $ECM(\widehat{\boldsymbol{\Sigma}})$, pero a su vez, este estimador en tales condiciones presenta la mejor respuesta en cuanto al sesgo absoluto de la estimación de $\sqrt{|\boldsymbol{\Sigma}|}$.

- Bajo factores de inflación δ_x y/o δ_y=1.5, se percibe que el procedimiento Fase I usando el estimador MCD presenta alguna ventaja frente a los demás estimadores en $ECM(\widehat{\boldsymbol{\Sigma}})$, cuando la probabilidad de contaminación es nula o moderada, (ϵ =0, 0.05), sin embargo, la ventaja que inicialmente alcanza desaparece en la medida que n y m se hacen cada vez más grandes

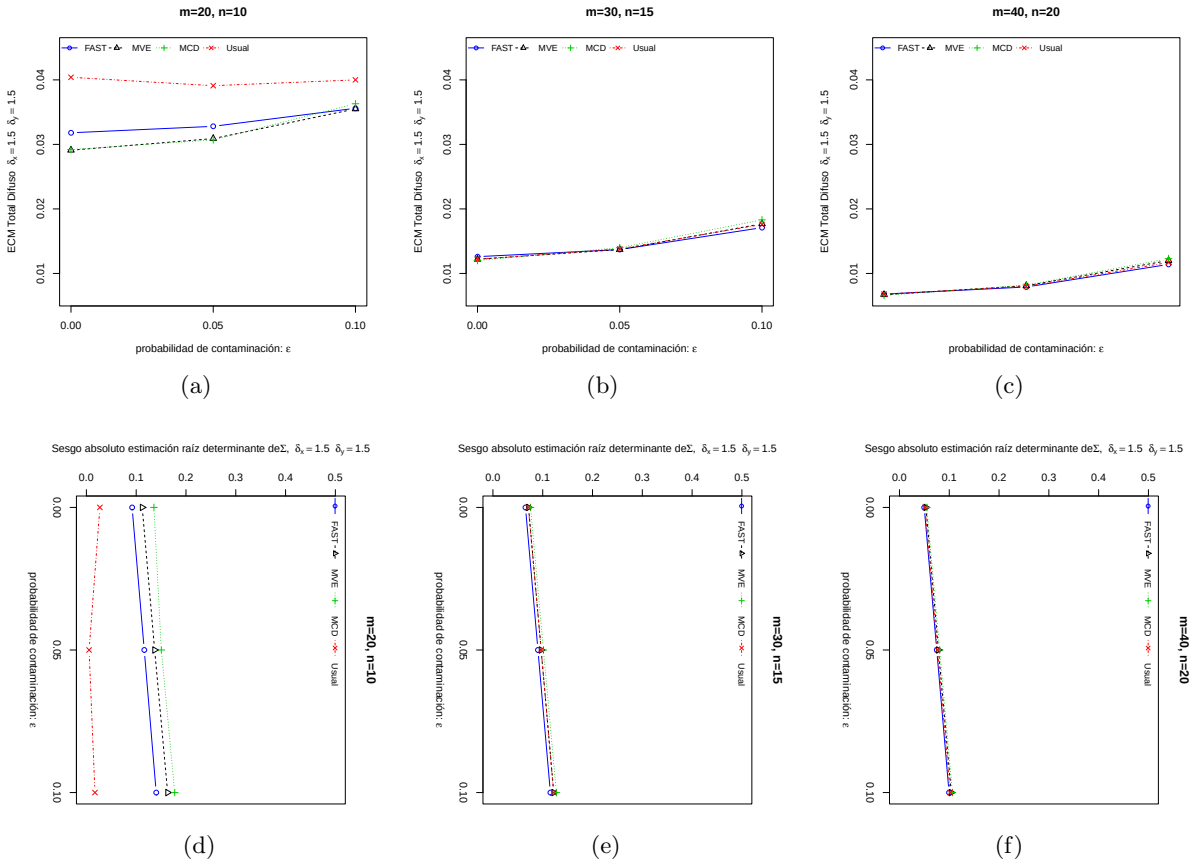

Figura 3: Resultados bajo contaminación difusa con δ_x = δ_y =1.5

y/o crece la probabilidad de contaminación. Con respecto al sesgo absoluto en la estimación de $\sqrt{|\mathbf{\Sigma}|}$, la respuesta de este estimador es peor comparada con la de los demás estimadores en todos los escenarios.

- Bajo condiciones de factor de inflación δ_x y/o δ_y=1.5, para el estimador S-FAST, se encontró que su desempeño en cuanto a $ECM(\widehat{\mathbf{\Sigma}})$ y al sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$ mejora a medida que n >10 y m >20, sin embargo, la ventaja que inicialmente alcanza desaparece en la medida que n y m se hacen cada vez más grandes y/o crece la probabilidad de contaminación.

- Cuando el factor de inflación de la varianza se elevó a δ_x y/o δ_y=3, bajo contaminación difusa en todos los escenarios se presenta peor el estimador MCD tanto en $ECM(\widehat{\mathbf{\Sigma}})$ como en el sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$. Por el contrario se percibe que usando el estimador S-FAST, muestra alguna ventaja bien diferenciada frente a los demás estimadores tanto en el $ECM(\widehat{\mathbf{\Sigma}})$ como en el sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$, cuando n >10 y m >20, ventaja que se mantiene en la medida que estos valores crecen y/o crece el número de

subgrupos contaminados.

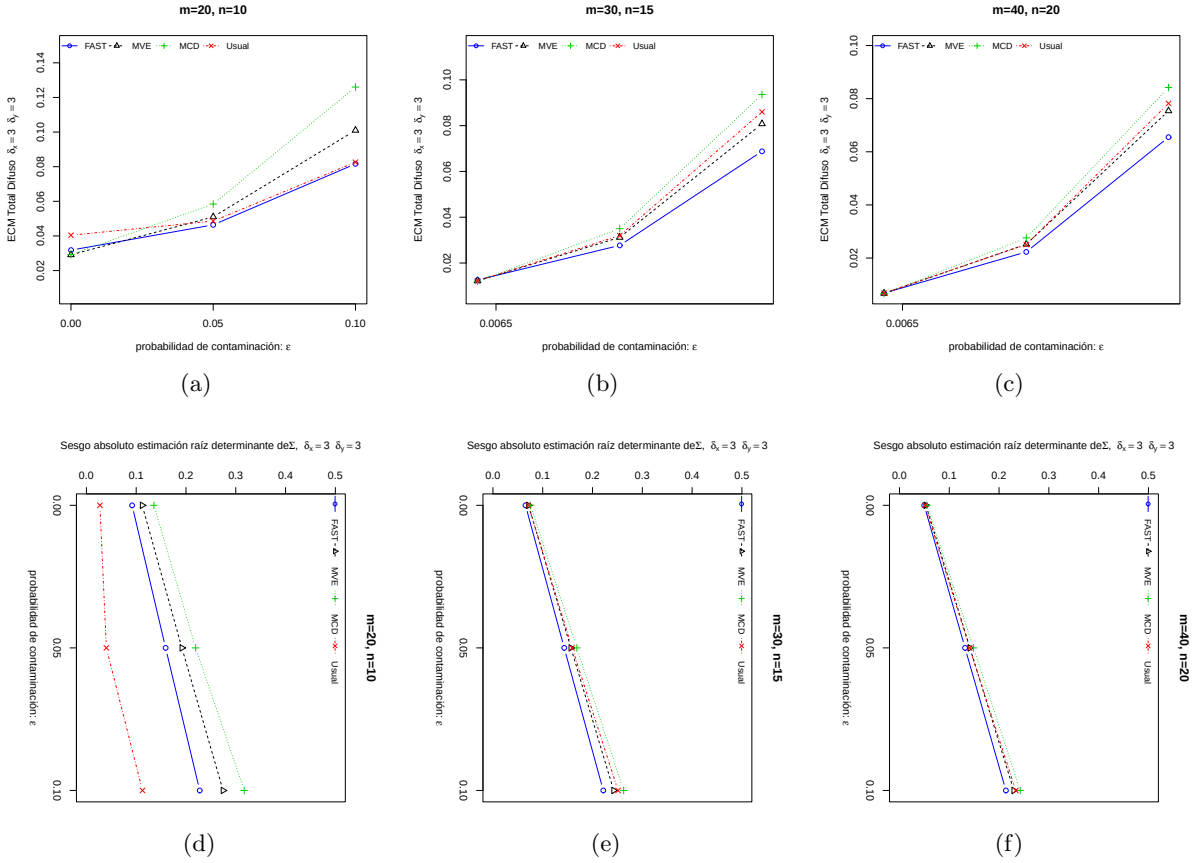

Figura 4: Resultados bajo contaminación difusa con δ_x = δ_y =3

- De acuerdo con los resultados obtenidos, Se podría asumir que a mayor probabilidad de contaminación, mayor tamaño de muestra y mayor número de subgrupos, en el caso de contaminación del tipo difuso, se presenta con mejor $ECM(\widehat{\mathbf{\Sigma}})$ y sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$ el estimador S-FAST y en niveles bajos de todos estos parámetros es mejor en $ECM(\widehat{\mathbf{\Sigma}})$ el estimador robusto MCD, sin embargo, se aprecia que la ventaja de los procedimientos basados en los estimadores robustos MVE y MCD respecto del estimador Usual no es muy alta y que para algunos escenarios se registró mejor el $ECM(\widehat{\mathbf{\Sigma}})$ bajo este último.

- En particular bajo condiciones de factor de inflación $\delta_x = \delta_y$=3 la respuesta favorece más al estimador S-FAST tanto en $ECM(\widehat{\mathbf{\Sigma}})$ como en el sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$.

Al evaluar el $ECM(\widehat{\mathbf{\Sigma}})$ y el sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$ de los estimadores Usual, MCD, MVE y S-FAST, con la carta de control $\sqrt{|\mathbf{S}|}$ sobre la muestra final de la Fase I y asumiendo

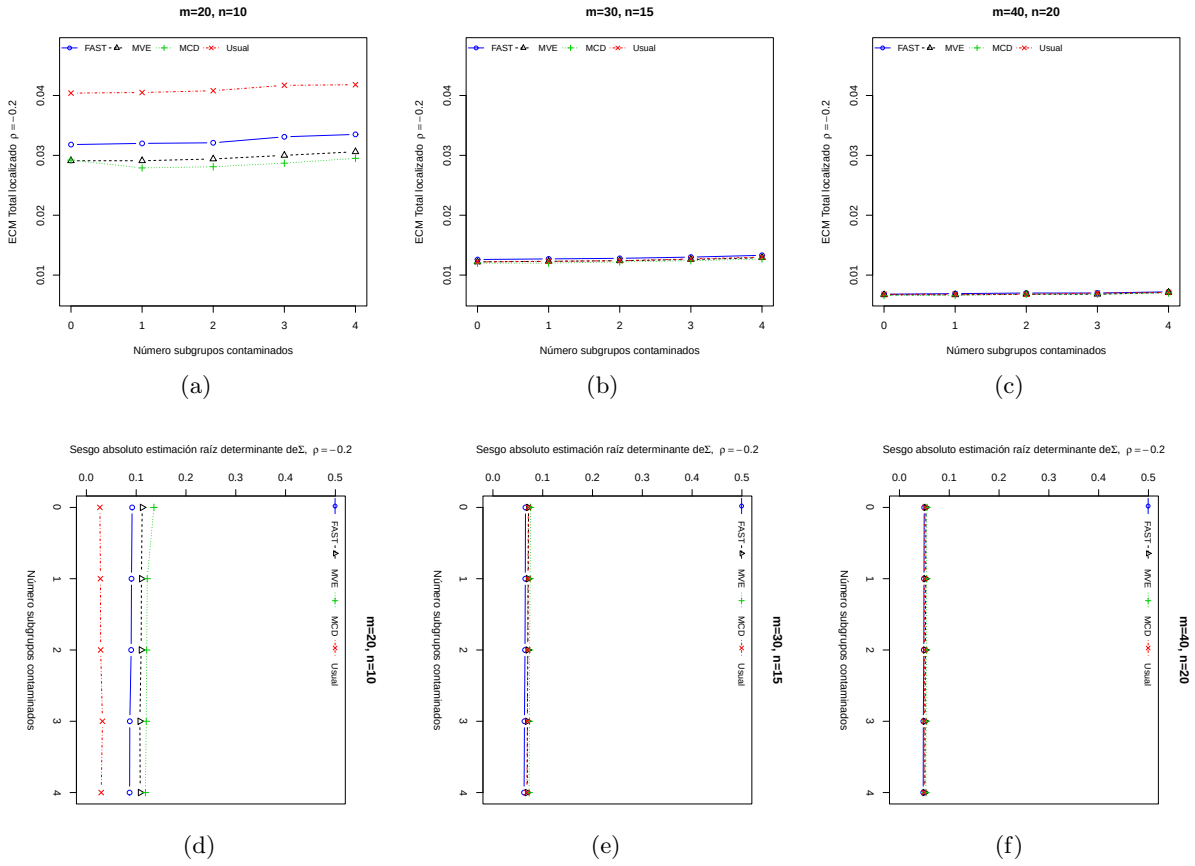

Figura 5: Resultados bajo contaminación localizada con $\rho = -0.2$

contaminación localizada afectando la correlación, se presentaron los siguientes hallazgos, (ver Figura 5 y 6)

- Se pudo establecer que a partir de los diferentes escenarios contemplados con el procedimiento Fase I al comparar los procesos con los diferentes estimadores utilizados, los mejores resultados en cuanto al $ECM(\widehat{\mathbf{\Sigma}})$ de la carta de control se presentan bajo el estimador MCD en todos los escenarios, sin embargo, la ganancia disminuye o se pierde a medida que el tamaño de muestra y el número de subgrupos crecen. Para tamaños de muestra pequeños n =10, en todos los escenarios se presentó peor y muy bien diferenciado el estimador usual, pero al incrementar el tamaño de la muestra n =15 y 20, en primer lugar, el peor desempeño en cuanto al $ECM(\widehat{\mathbf{\Sigma}})$ de la carta se encontró con el estimador S-FAST y en segundo lugar, el $ECM(\widehat{\mathbf{\Sigma}})$ con el estimador Usual se hizo más próximo al que se logró con los estimadores MVE y MCD.

- Con respecto al sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$, la situación es contraria, pues para

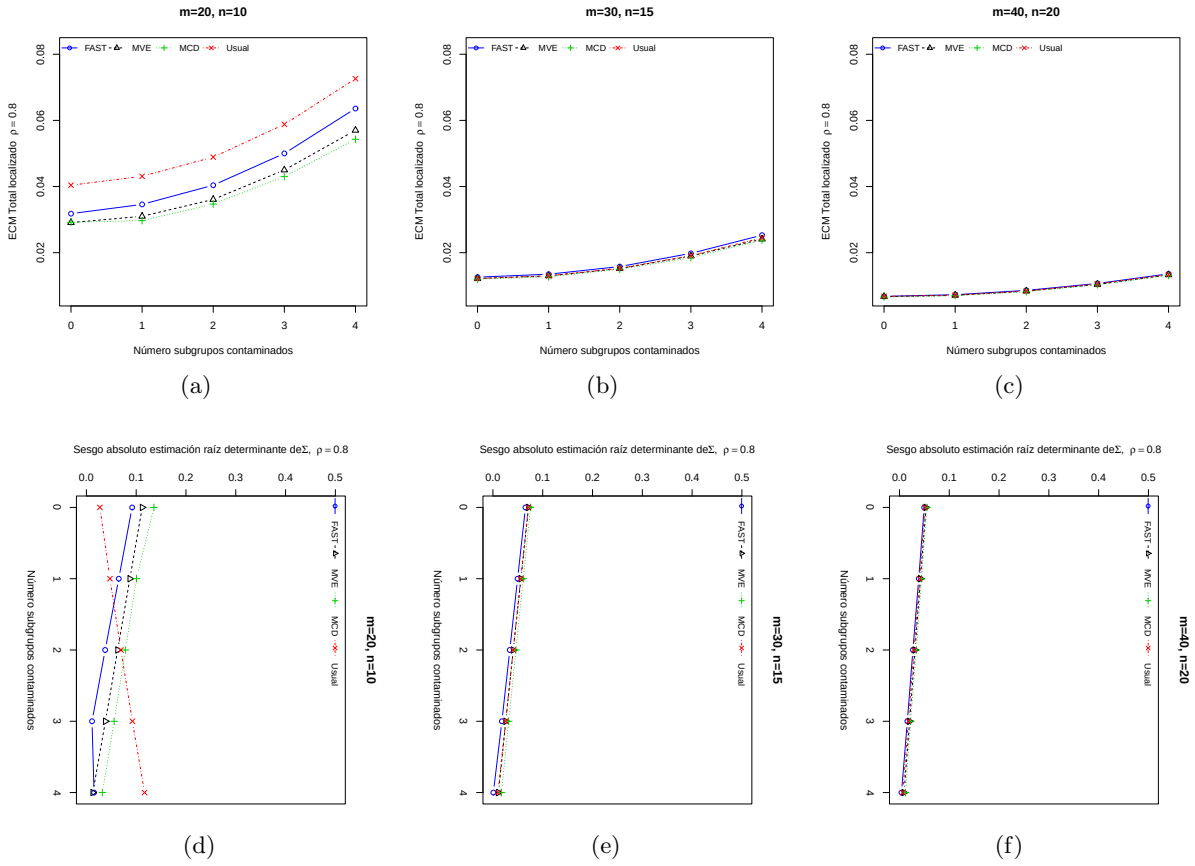

Figura 6: Resultados bajo contaminación localizada con ρ =0.8

el estimador MCD se registra los peores resultados mientras que para el estimador S-FAST el sesgo absoluto en todos los casos es mejor.

5. CONCLUSIONES

- En todos los escenarios de contaminación considerados, con n=10, el procedimiento Fase I usando el estimador Usual, se presenta la peor condición en cuanto al $ECM(\widehat{\mathbf{\Sigma}})$, pero a su vez, este estimador en tales condiciones presenta la mejor respuesta en cuanto al sesgo absoluto de la estimación de $\sqrt{|\mathbf{\Sigma}|}$.

- En general, con base en los resultados que aportan los procesos de simulación y asumiendo contaminación localizada, en los que se implementó factor de inflación moderado (δ_x y/o δ_y=1.5) y alto, (δ_x y/o δ_y=3), El procedimiento Fase I diferencia claramente al estimador MCD con alguna ventaja frente a los demás estimadores robustos y el estimador Usual, al compararse en cuanto al $ECM(\widehat{\mathbf{\Sigma}})$, cuando el número de subgrupos contaminantes es bajo,

$k \leq 2$, tamaño de muestra pequeño, n =10, y el número de subgrupos m =20; ventaja que disminuye o desaparece al incrementar el tamaño de muestra, el número de subgrupos y/o el número de subgrupos contaminados, sin embargo, en cuanto al sesgo absoluto de la estimación de $\sqrt{|\boldsymbol{\Sigma}|}$, presenta la peor condición en todos los escenarios observados.

- En los mismos términos de la conclusión anterior, con respecto al estimador S-FAST se encontró que su desempeño en cuanto al $ECM(\widehat{\boldsymbol{\Sigma}})$ y sesgo absoluto de la estimación de $\sqrt{|\boldsymbol{\Sigma}|}$, mejora a medida que n >10 y m >20, sin embargo, al ser comparado con los demás estimadores, la ventaja que inicialmente alcanza desaparece en la medida que n y m se hacen cada vez más grandes y/o crece el número de subgrupos contaminados.

- Bajo contaminación difusa perturbando solo varianzas, de acuerdo con los resultados obtenidos, Se podría asumir que a mayor probabilidad de contaminación, mayor tamaño de muestra y mayor número de subgrupos, se presenta con mejor ventaja en cuanto al $ECM(\widehat{\boldsymbol{\Sigma}})$ y sesgo absoluto de la estimación de $\sqrt{|\boldsymbol{\Sigma}|}$, el estimador S-FAST y en niveles bajos de todos estos parámetros es mejor en precisión el estimador robusto MCD, sin embargo, se aprecia que la ventaja de los procedimientos basados en el estimador robusto MCD respecto de los estimadores MVE y Usual no es muy alta y que para algunos escenarios se registró mejor el $ECM(\widehat{\boldsymbol{\Sigma}})$ bajo el estimador Usual, razón por la que se podría concluir que bajo estos términos, no se obtienen resultados positivos en cuanto al $ECM(\widehat{\boldsymbol{\Sigma}})$ de los estimadores para los casos robustos MCD y MVE comparados con el procedimiento basado en el estimador Usual.

- Bajo contaminación localizada afectando la correlación se pudo establecer que a partir de los diferentes escenarios tenidos en cuenta con el procedimiento Fase I, al comparar los procesos con los diferentes estimadores utilizados, los mejores resultados en cuanto al $ECM(\widehat{\boldsymbol{\Sigma}})$ de la carta de control se presentan bajo el estimador MCD en todos los escenarios, sin embargo, la ganancia disminuye o se pierde a medida que el tamaño de muestra y el número de subgrupos crecen. Con respecto al sesgo absoluto de la estimación de $\sqrt{|\boldsymbol{\Sigma}|}$, la situación es contraria, pues para el estimador MCD se registra los peores resultados mientras que para el estimador S-FAST en todos los casos es el mejor.

- Teniendo en cuenta que los estimadores robustos presentan mejores condiciones en cuanto a $ECM(\widehat{\boldsymbol{\Sigma}})$ que los estimadores clásicos, llama la atención que se produzcan respuestas en las que la estimación con base en el estimador clásico se muestre mejor en el $ECM(\widehat{\boldsymbol{\Sigma}})$ que con los estimadores robustos; solo se logra una leve mejoría en algunos casos. Queda la pregunta, ¿será posible que para el proceso resulte muy crítica cualquier desviación de la matriz de covarianzas?

- Es importante señalar el hecho que todos los resultados obtenidos están en dependencia con

la forma en que han sido implementados los estimadores en R, es decir, el tratamiento dado en el proceso de construcción de estimadores robustos de Σ_0 combinando la información de m muestras independientes. En esta investigación, se optó por utilizar el promedio de las matrices de los m subgrupos de muestras independientes de tamaño n; sin embargo, se podría utilizar alternativamente el método propuesto por He & Fung (2000) para estimadores S, que fue adaptado más tarde por Hubert y Van Driessen (2004) para estimadores MCD. Bajo este método en lugar de agrupar las matrices de covarianzas de los subgrupos, se agrupan las observaciones y se obtiene una única muestra y sobre esa gran muestra finalmente se aplica la estimación robusta para obtener el estimador robusto de la matriz de covarianzas.

Referencias

Alt, F. B. (1985). Multivariate quality control. The Encyclopedia of Statistical Sciences, Kotz S, Johnson NL, Read CR (eds.), Wiley: New York, 110–122.

Alt, F. B. & Smith, N. D. (1988). Multivariate process control. Handbook of Statistics, Elsevier: Amsterdam. 333–351.

Anderson, T. W. (1984). An introduction to multivariate statistical analysis. Second edition, John Wiley and Sons, Inc., New York. 264.

Barnett, V. (1979). Some outlier test for multivariate samples, *South African Statistical Journal*, 13, 29–52.

Barnett, V. & Lewis, T (1998). Outliers in statistical data. Third ed. John Wiley & Sons, Inc: New York.

Butler R. W., Davies, P. L. & Jhun M. (1993). Asympthotic for the minimum covariance determinant estimator. *The Annals of Statistics*, 21, 1385–1401.

Cerioli, A., Riani, M. C. & Atkinson, A. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter, *Stat Comput.* 19, 341–353.

Chenouri, S. E., Variyath, A. M. & Steiner. S. H. (2009). A multivariate robust control chart for individual observations, *Journal of Quality Technology*, 41(**3**), 259–271.

Croux, C. & Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71, 161–190.

Chenouri, S. & Variyath, A. M. (2010). A comparative study of Phase II robust multivariate control charts for individual observations, *QREI*, 27, 857–865.

Davies, P. L. (1987). Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15, 1269–1292.

Davis, W. S. & Yen, D. C. (1998). Handbook. The information Systems: analysis and design. Disponible en: [$http://www.hit.ac.il/staff/leonidm/information-systems/ewtoc.html$.]

Djauhari, M. A., Mashuri, M. & Herwindiati, D. E. (2008). Multivariate process variability monitoring. *Communication in Statistics - Theory and Methods*, 37, 1742–1754.

Gomes, J. H. F., Paiva, A. P., Costa, S. C., Balestrassi, P. P. & Paiva, E. J. (2012). Weighted multivariate mean square error for processes optimization: A case study on flux-cored arc welding for stainless steel claddings. *European Journal of Operational Research*, 226(**2013**), 522–535.

He, X. & Fung W. K. (2000). High breakdown estimation for multiple populations with applicatons to discriminant analysis. *Journal of Multivariate Analysis*, 72, 151–162.

Hubert, M., Rousseeuw, P. J. & van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science. Institute of Mathematical Statistics*, 23(**1**), 92–119.

Jensen, W. A., Birch, J. B. & Woodall W. H. (2006). High breakdown estimation methods for phase I multivariate control charts. Technical Report 05-6. Disponible en: [$http://www.web-e.stat.vt.edu/dept/web-e/tech_reports/TechReport05-6.pdf$].

Jones-Farmer, L. A., Woodall, W. H., Steiner, S. H. & Champ, C. W. (2014). An overview of phase I analysis for process improvement and monitoring. *Journal of Quality Technology*, 46(**3**), 265.

Köksoy, O. (2006). Multiresponse robust design: Mean square error (MSE) criterion. *Applied Mathematics and Computation*, 175, 1716–1729.

Lopuhaä, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics*, 17, 1662–1683.

Lopuhaä, H. P. & Rousseeuw, P. J. (1991). Breakdown points of affine eqivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19 (**1991**), 229–248.

Maronna, R. A., Martin, D. & Yohai, V. (2006). Robust statistic, ISBN: 978-0-470-01092-1, Wiley Series.

Montgomery, D. (2005). Control estadístico de la calidad, Tercera edición, México: Limusa Wiley.

Nelson, L. S. (1988). Control Charts: Rational Subgroups and Effective Applications. *Journal of Quality Technology*, 20(**1**), 73–75.

Pison, G., Van Aelst, S. & Willems, G. (2002). Small Sample Corrections for LTS and MCD. *Metrika*, 55, 111–123.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponible en: http://www.R-project.org/.

Rocke, D. M. & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 89, 888–896.

Rousseeuw, P. J. & Leroy A. M. (1987). Robust regression and outlier detection. John Wiley and Sons, Inc. New York.

Rousseeuw, P. J. & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.

Rousseeuw, P. J. & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633–639.

Rousseeuw, P. J. & Yohai, V. J. (1984). Robust regression by means of S-estimators. *Robust and Nonlinear Time Series (Lecture, Notes in Statistics No 26)*, cds, J. Frankc, W, Härdle, and R.D. Martin, New York: Springer-Verlag, 256–272.

Salibian-Barrera, M. & Yohai, V. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, 15, 414–427.

Schoonhoven, M., Nazir, H. Z., Riaz, M. & Does, R. J. M. M.(2011). Robust location estimaters for the $\bar{X}$ control chart. *Journal of Quality Technology*, 43(**4**), 363–379.

Schoonhoven, M. & Does, Ronald J. M. M. (2012). A robust standard deviation control chart. *Technometrics*, 54(**1**), 73–82.

Tatum, L. G. (1997). Robust estimation of the process standard deviation for control charts. *Technometrics*, 39(**2**), 127–141.

Todorov, V. (2008). A note on the MCD consistency and small sample correction factors. Unpublished manuscript, in preparation.

Todorov, V. & Filzmore, P. (2009). An object oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(**3**).

Todorov, V. (2012). Scalable robust estimators with high breakdown point. Disponible en http://cran.r-project.org/web/packages/rrcov/rrcov.pdf. Consultado 15-03-2013.

Vargas, J. A. (2003). Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology*, 35, 367–376.

Vargas, J. A. & Lagos, J. (2007). Comparison of multivariate control charts for process dispersion. *Quality Engineering*, 19, 191–196.

Variyath, A. M. & Vattathoor, J. (2013). Robust control charts for monitoring process variability in Phase I multivariate individual observations. *Journal of Quality and Reliability Engineering*, 30 (**6**), 795–812.

Woodall, W. H. (2000). Controversies and xontradictions in statistical process control. *Journal of Quality Technology*, 32(**4**), 341–350.

Yañez, S., González, N. & Vargas, J. A. (2010). Hotellingt's T2 control charts based on robust estimators. *Dyna*, 163, 239–247.