**Revista Colombiana de Estadística**
Volumen 26 N° 2. Págs. 159 *a* 179. Diciembre 2002

# Carta $T^2$ con base en estimadores robustos de los parámetros

Sergio Yáñez C.[*]
José A. Vargas[**]
Nelfi González A.[***]

**Resumen**

En la primera etapa de implementación de un sistema de control multivariado, usando la carta $T^2$ de Hotelling con $n$ observaciones históricas individuales, la presencia de *outliers* distorsiona la estimación de los parámetros del proceso y del límite de control debido al efecto de enmascaramiento. En este trabajo proponemos el uso de estimadores robustos para la construcción del estadístico $T^2$ en esta primera etapa. Se prueba con estimadores MVE (elipsoide de mínimo volumen) y estimadores S biponderados, para el caso $p = 2$. Los resultados de simulaciones señalan que estos dos procedimientos resultan consistentes en la detección de *outliers* provenientes de perturbaciones en el vector de medias y de la matriz de varianzas covarianzas, consideradas individual y conjuntamente, con diferentes niveles de contaminación.

***Palabras clave***: *Control multivariado, $T^2$ de Hotelling, estimadores robustos, outliers multivariados, enmascaramiento.*

Investigación patrocinada por la Dirección de Investigación, Universidad Nacional de Colombia, Sede Medellín (DIME). Proyecto código 2010100822.

[*]Profesor asociado, Escuela de Estadística, Universidad Nacional de Colombia, Sede Medellín. E-mail: syanez@perseus.unalmed.edu.co

[**]Profesor asociado, Departamento de Estadística, Universidad Nacional de Colombia, Sede Bogotá. E-mail: avargas@matematicas.unal.edu.co

[***]Instructora asociada, Escuela de Estadística, Universidad Nacional de Colombia, Sede Medellín. E-mail: ngonzale@perseus.unalmed.edu.co

**Abstract**

In Phase I, Stage 1 of a multivariate process control, the implementation of a Hotelling's $T^2$ chart with $n$ individual observation, *outliers* cause difficulties with the estimation of process parameters and control limits due to masking effects. We propose procedures to construct robust estimators based upon the MVE (Minimum Volume Ellipsoide) and the biweighted S estimator, for case $p = 2$ (Bivariate Process). Simulation results show the good perfomance of these estimators before *outliers* presence, avoiding masking effects, when we are estimating the mean vector and varianza covarianza matrix, both individually and jontly. We make the investigation with different levels of contamination affecting the mean vector and varianza covarianza matrix.

***Key words****: Multivariate control, Hotelling's $T^2$ chart, robust estimator, masking outliers.*

# 1. Introducción

El control estadístico multivariado consiste en el monitoreo simultáneo de dos o más características de calidad, y para este fin suele emplearse la carta $T^2$ de Hotelling bajo la presunción de normalidad multivariada. Alt & Smith (1988) definen un proceso de dos fases para el establecimiento del sistema de control: fase I y fase II. La fase I es desarrollada en dos etapas denominadas etapa 1 y etapa 2; en la primera los parámetros del proceso, $\mu_{\mathbf{0}}$ y $\mathbf{\Sigma_0}$, son estimados a partir de datos históricos; considerando el caso cuando éstos representan $n$ observaciones individuales $\mathbf{x}_i$ del proceso $p$ variado, típicamente se grafica el estadístico:

$$T_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}), \tag{1}$$

que se distribuye como

$$\frac{(n-1)^2}{n} B_{p/2,(n-p-1)/2}, \tag{2}$$

(Tracy, Young & Mason 1992), donde $B_{p/2,(n-p-1)/2}$ representa la distribución beta con parámetros $p/2$ y $(n-p-1)/2$,

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i,$$

y

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T,$$

son los respectivos estimadores muestrales usuales insesgados de los parámetros desconocidos $\mu_0$ y $\mathbf{\Sigma}_0$ que definen al proceso bajo control. Un límite de control UCL igual a

$$\frac{(n-1)^2}{n} B_{\alpha, p/2, (n-p-1)/2} \tag{3}$$

es fijado con las observaciones que se consideran fueron tomadas bajo control. En la etapa 2 se usan las estimaciones halladas en la etapa 1 y se verifica con nuevas observaciones si el proceso continúa siendo estable. Finalmente en la fase II, los estimadores obtenidos en la fase I son usados como parámetros definitivos para el proceso.

Según Sullivan & Woodall (1996), la presencia de $N$ *outliers* dispersos aleatoriamente entre las $n$ observaciones de la etapa 1 constituye una causa de fuera de control; la carta $T^2$ usual carece de robustez ante la presencia de tales observaciones, por un fenómeno conocido como enmascaramiento, el cual inhibe al procedimento para detectar cualquier señal.

Se han planteado algunas soluciones, por ejemplo, Atkinson & Mulira –citados por Sullivan & Woodall (1996)– desarrollaron un método gráfico denominado *Stalactite Chart*, pero su uso requiere destreza para la interpretación. Sullivan & Woodall (1996) proponen una modificación al anterior método; sin embargo resulta poco efectivo ante desviaciones escalonadas en el vector de medias ubicadas en la mitad de los datos. Vargas (2002) explora soluciones mediante el uso de estimadores robustos para el vector de medias y la matriz de varianzas covarianzas; específicamente trata con estimadores MVE (*Elipsoide de un estimador "recortado" aplicado sobre las distancias de Mahalanobis*). Evalúa en el caso bivariado el comportamiento de las cartas robustas construidas, frente a contaminaciones arbitrarias con observaciones en las que se ha producido una desviación en la media, llegando a la conclusión de que la carta $T^2$ construida en la etapa 1 mediante estimadores MVE es una buena alternativa para la detección de *outliers*.

En el presente trabajo se prueba de nuevo con la carta $T^2$ basada en el MVE y se compara con otra versión robusta obtenida usando estimadores $S$[1],

[1]También se probó con estimadores Stahel-Donoho; sin embargo sólo exhibieron ventajas en casos muy particulares. Éstos son los primeros estimadores afin equivariantes de localización y dispersión multivariada, con un punto de ruptura de 0.5 desarrollados independientemente por Stahel (1981) y Donoho (1982). Véanse Maronna & Yohai (1995), Hampel, Ronchetti, Rousseeuw & Stahel (1986), Becker & Gather (1999).

frente a la presencia de *outliers* provenientes de tres esquemas de contaminación: desviaciones en el vector de medias, inflación de la matriz de varianzas covarianzas, y combinación de los dos anteriores tipos de perturbación, lo que denominaremos contaminación cruzada.

## 2. Estimadores robustos

Rousseeuw y Yohai (1984) –citados por Rousseeuw & Leroy (1987)– introdujeron en el campo de regresión la clase de estimadores $S$. Una generalización posterior de estos estimadores para localización multivariada y covarianza fue hecha definiendo los estimadores $S$ de localización y de forma multivariados, como el vector $\mathbf{t}$ y la matriz semidefinida positiva $\mathbf{C}$, tales que:

$$\text{minimizan a } |k^2\mathbf{C}|, \tag{4}$$

sujetos a:

$$n^{-1}\sum \rho([(\mathbf{x_i}-\mathbf{t})^T(k^2\mathbf{C})^{-1}(\mathbf{x_i}-\mathbf{t})]^{1/2}) = b_0, \tag{5}$$

o bien,

$$n^{-1}\sum \rho(d_i/k) = b_0,$$

(Woodruff & Rocke 1994), en donde $\rho$ corresponde a la función biponderada[2] dada por:

$$\rho_b(d;c) = \begin{cases} d^2/2 - d^4/(2c^2) + d^6/(6c^4), & 0 \le d \le c, \\ c^2/6, & d > c, \end{cases}$$

donde las constantes $c$ y $b_0$ son ajustadas para alcanzar un punto de ruptura de 0.5, que de acuerdo con Rousseeuw & Leroy (1987, p. 142, 1987) para $p = 2$ corresponden a 1.547 y 0.1995 respectivamente, con $b_0 = (1/2)\rho(c)$.

Estos estimadores poseen "alto punto de ruptura", es decir, cercano a 0.5, lo cual significa que el estimador de localización multivariado permanece limitado y que los valores propios del estimador de la matriz de covarianza son lejanos de 0 y de $\infty$ cuando menos de la mitad de los datos son remplazados por valores arbitrarios (Rousseeuw & Zomeren 1990). También son afin equivariantes, lo que implica que si $\mathbf{C}$ son respectivamente los estimadores de localización y covarianzas, basados en $\mathbf{x_1}, \ldots, \mathbf{x_n}$, con $\mathbf{x_i} \in R^p$, entonces, para cualquier vector $\mathbf{b}$ y cualquier matriz no singular $\mathbf{A}$ (Rousseeuw & Leroy 1987):

$$\mathbf{t}(\mathbf{Ax_1}+\mathbf{b}, \ldots, \mathbf{Ax_n}+\mathbf{b}) = \mathbf{At}(\mathbf{x_1}, \ldots, \mathbf{x_n}) + \mathbf{b}, \tag{6}$$

[2]Se usa esta función según resultados alcanzados por Woodruff & Rocke (1994) y Rocke & Woodruff (1996) en sus trabajos sobre identificación de *outliers* y estimación robusta.

$$\mathbf{C}(\mathbf{Ax_1} + \mathbf{b}, \ldots, \mathbf{Ax_n} + \mathbf{b} = \mathbf{At}(\mathbf{x_1}, \ldots, \mathbf{x_n})\mathbf{A}^T. \tag{7}$$

La definición de un estimador $S$ permite aplicar un esquema iterativo para aproximar su cálculo. Woodruff & Rocke (1994) proponen un método basado en la definición del estimador S como el mínimo global de la función objetivo, el cual arroja un par $t^{(j)}$ y $C^{(j)}$ (los estimadores resultantes en $j$ pasos del proceso iterativo). Adicionalmente, para obtener los estimadores definitivos se realiza una reponderación, calculando las distancias de Mahalanobis $d_i^2$ con $t^{(j)}$ y $C^{(j)}$ obtenidos en el esquema anterior, las cuales se comparan con $u = (1 + 15/(n-p))^2 \chi^2_{0.95} \text{med}\{d_i^2\}/\chi^2_{0.5}$ (Maronna & Yohai 1995), donde $\chi^2_{0.95}$ y $\chi^2_{0.5}$ son los cuantiles 95 y 50 de la distribución chi cuadrado con $p$ grados de libertad, y $\text{med}\{d_i^2\}$ es la mediana de las distancias de Mahalanobis. A las observaciones con $d_i^2 \leq u$ se les asigna un peso $w_i$ de 1; en caso contrario el peso es cero. Los estimadores finales son calculados como:

$$T = \left( \sum_i^n w_i \right)^{-1} \sum_{i=1}^n w_i \mathbf{x}_i, \tag{8}$$

$$C = \left( \sum_i^n w_i - 1 \right)^{-1} \sum_{i=1}^n w_i (\mathbf{x}_i - T)(\mathbf{x}_i - T)^t. \tag{9}$$

En experimentos de simulación, Woodruff & Rocke (1994) sugieren que en algoritmos para estimación en dos etapas, sea usado el MCD para los estimadores iniciales en el esquema iterativo para la búsqueda aproximada de los estimadores $S$.

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 MVE de covarianza, $\mathbf{C}$, corresponde al volumen de dicho elipsoide multiplicado por un factor de corrección para obtener consistencia (Rousseeuw & Leroy 1987, pág. 258).

Más formalmente (Rousseeuw & Zomeren 1990): el estimador MVE de localización y covarianza multivariado es el par $(\mathbf{t}, \mathbf{C})$ tales que el determinante de $\mathbf{C}$ es minimizado sujeto a:

$$\#\{i; (\mathbf{x_i} - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x_i} - \mathbf{t}) \leq a^2\} \geq h, \tag{10}$$

donde

$$h = \left[ \frac{n + p + 1}{2} \right], \tag{11}$$

con $[k]$ la parte entera de $k$; $a^2$ 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. Cuando el tamaño de muestra es pequeño puede requerirse un factor de corrección $c^2_{n,p}$ que depende de $n$ y $p$. Rousseeuw & Zomeren (1990) sugieren usar $c^2_{n,p} = \big(1 + 15/(n-p)\big)^2$ como un factor de corrección razonable para muestras pequeñas.

Una justificación intuitiva de este método, como lo expresa Peña (2002), consiste en la idea de que los puntos atípicos estarán en los extremos de la distribución, por lo que se puede buscar una zona de alta concentración de puntos que presumiblemente serán puntos buenos, y con los cuales se determine el centro de los datos y la matriz de covarianzas. Para hallar esta zona de alta densidad de puntos se exige que el elipsoide que cubra al menos el 50% de los datos tenga volumen mínimo.

En cuanto al cálculo de este estimador, se tiene que en muchos casos no es factible considerar todas las "mitades" de los datos para calcular el volumen del elipsoide más pequeño alrededor de ellos; así que algoritmos basados en el remuestreo han sido implementados para el cálculo aproximado. Otra opción en particular, la cual es usada en este trabajo, es la que aplica la rutina SPLUS *cov.mve*, basada en un algoritmo genético.

## 3. Construcción de la carta $T^2$ robusta

Dada una muestra aleatoria de $n$ observaciones históricas individuales de un proceso normal bivariado ($p = 2$), el estadístico $T^2$ de la carta robusta para la etapa 1 se obtiene mediante la sustitución de los estimaciones usuales insesgadas del vector de medias y de la matriz de covarianzas por las estimaciones correspondientes al procedimiento de estimación robusto considerado; para el caso, las estimaciones derivadas de $T^2_{r,i} = (\mathbf{x_i} - \mu_\mathbf{r})^\mathbf{T}\mathbf{\Sigma_r^{-1}}(\mathbf{x_i} - \mu_\mathbf{r})$, $T^2_{r,i} = (\mathbf{x_i} - \mu_\mathbf{r})^\mathbf{T}\mathbf{\Sigma_r^{-1}}(\mathbf{x_i} - \mu_\mathbf{r})$, donde $\mu_\mathbf{r}$ y $\mathbf{\Sigma_r^{-1}}$ son las correspondientes estimaciones robustas del vector de medias y la matriz de covarianza. Los UCL, tanto de la carta usual como para las versiones robustas, son determinados mediante simulación fijando una tasa de falsa alarma total de 0.05; ésta es una tasa total dado que los correspondientes $T^2_{r,i}$ de las $n$ observaciones históricas del proceso son comparados sigue el mismo procedimiento empleado por Sullivan & Woodall (1996); sin pérdida de generalidad (por las propiedades de invarianza del estadístico $T^2$) se considera que los parámetros del proceso en control son $\mathbf{\Sigma} = \mathbf{I_2}$ y $\mu_\mathbf{0} = \mathbf{0}$, y mediante 5000 muestras aleatorias de tamaño $n = 30$, se halla el UCL como el percentil 95 de la distribución de los máximos de $T^2_{r,\ i}$.

Considere la siguiente notación para distinguir las cartas robustas:

- **USUAL:** Carta $T^2$ obtenida mediante estimadores usuales.
- **MVE:** Carta $T^2$ obtenida mediante estimadores MVE.
- **SE:** Carta $T^2$ obtenida mediante estimadores $S$ biponderados - método iterativo.

Siguiendo el procedimiento descrito, se obtuvieron los valores en la tabla 1:

Tabla 1: UCL.

| **USUAL** | **MVE** | **SE** |
|---|---|---|
| 10.51234 | 24.93355 | 20.24410 |

# 4. Escenarios de simulación

Se prueban los tres procedimientos mediante la simulación de 1000 muestras de tamaño $n = 30$, contaminadas con $N = 1, 2, 3, 4, 5, 6, 7$ *outliers* dispersos aleatoriamente, provenientes de los escenarios descritos en la tabla 2.

Tabla 2: Escenarios de contaminación.

| **Tipo de contaminación** | **Valor parámetros** | **Total escenarios** |
|---|---|---|
| Desviación de la media<br>$N_p(\delta, \mathbf{I}_2)$ | No centralidad<br>$d^2 = \parallel \delta \parallel^2 = 5, 10, 15, 20, 25$ | 35 |
| Contaminación simétrica<br>$N_p(\mathbf{0}, \lambda\mathbf{I}_2)$ | Factor de inflación<br>$\lambda = 1.5, 2, 2.5, 3.5, 4.5,$<br>$8, 10, 12, 16$ | 63 |
| Contaminación cruzada<br>$N_p(\delta, \lambda\mathbf{I}_2)$ | No centralidad<br>$d^2 = 5, 10, 15, 20, 25$<br>Factor de inflación<br>$\lambda = 1.5, 4.5, 8.5, 12.5$ | 140 |

# 5. Medidas de comparación

Sea:

- *NSIM:* Número de simulaciones, que corresponde a 1000.
- *N:* Número de *outliers* en cada muestra.
- *n:* Tamaño de muestra, corresponde a 30.

Las medidas de desempeño fueron definidas en forma similar a las definidas por Kosinski (1999):

- *Proporción promedio de* outliers *detectados:* Es la proporción promedio de *outliers* que en una muestra de tamaño 30 son detectados por un procedimiento, cuando hay $N$ total

$$\mathbf{pod}(N) = \frac{1}{NSIM} \sum_{i=1}^{NSIM} \left[ \frac{1}{N} \sum_{j=1}^{N} I(o_j = 1) \right], \tag{12}$$

  $o_j = 1$ indica que $T^{*2}_{r,j} \geq UCL$, donde $T^{*2}_{r,j}$ es la distancia cuadrada de Mahalanobis calculada para el $j$-ésimo *outlier* en la muestra de tamaño 30. $I(o_j = 1) = 1$, si se cumple la anterior condición.

- *Proporción promedio de enmascaramiento:* Es la proporción promedio de *outliers* enmascarados en muestras de tamaño 30 cuando hay $N$ *outliers.*

$$\mathbf{pen}(N) = \frac{1}{NSIM} \sum_{i=1}^{NSIM} \left[ \frac{1}{N} \sum_{j=1}^{N} I(m_j = 1) \right], \tag{13}$$

  $m_j = 1$ indica que $T^{*2}_{r,j} < UCL$, donde $T^{*2}_{r,j}$ es la distancia cuadrada de Mahalanobis calculada para el $j$-ésimo *outlier* en la muestra de tamaño 30. $I(m_j = 1) = 1$, si se cumple la anterior condición.

- *Proporción promedio de señales:* Es la proporción promedio de señales producidas por un procedimiento en muestras de tamaño 30, cuando hay $N$ *outliers.*

$$\mathbf{pse}(N) = \frac{1}{NSIM} \sum_{i=1}^{NSIM} \left[ \frac{1}{n} \sum_{j=1}^{n} I(l_j = 1) \right], \tag{14}$$

donde $l_j = 1$ indica que $T^2_{r,j} \geq UCL$. Luego $I(l_j = 1) = 1$ si se cumple la anterior condición.

- *Proporción promedio de* swamping*:* Es la proporción promedio de puntos no contaminantes que un procedimiento señala arriba de su UCL, en la presencia de $N$ *outliers* en la muestra.

$$\mathbf{psw}(N) = \frac{1}{NSIM} \sum_{i=1}^{NSIM} \left[ \frac{1}{n-N} \sum_{j=1}^{n-N} I(s_j = 1) \right], \qquad (15)$$

$s_j = 1$ indica que $T^{**2}_{r,j} \geq UCL$, donde $T^{**2}_{r,j}$ es la distancia cuadrada de Mahalanobis calculada para la $j$-ésima observación no contaminante en la muestra de tamaño 30. $I(s_j = 1) = 1$ si se cumple la anterior condición.

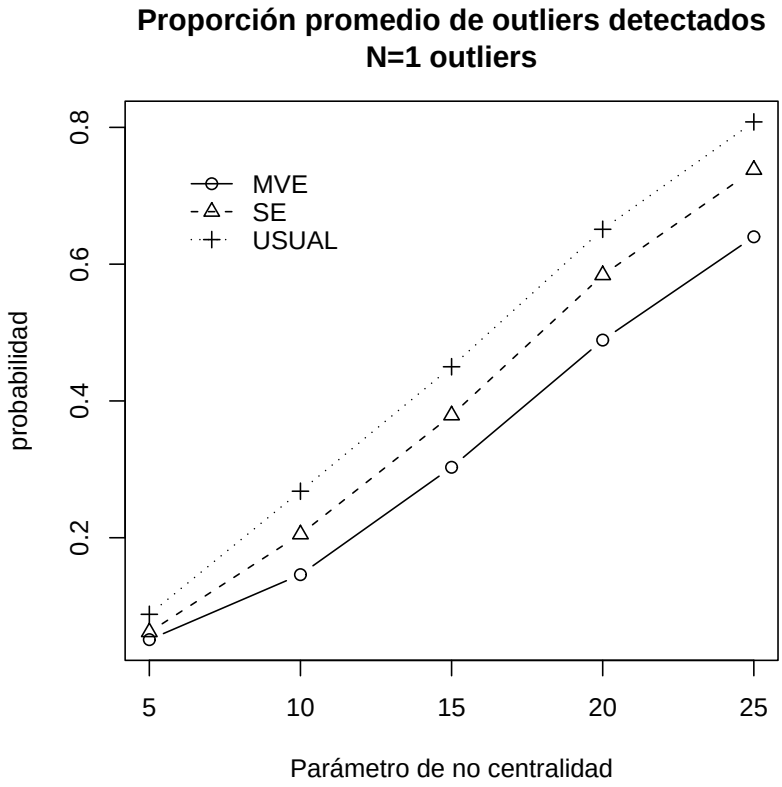
De estas cuatro medidas, la referente al *swamping* resultó poco informativa, en tanto que el promedio de detección de *outliers* permitió distinguir claramente las respuestas de cada procedimiento ante los diferentes niveles y tipos de contaminación. La proporción promedio de señales brindó información similar que la primera medida, sólo que la escala de respuesta es menor. Por otra parte el enmascaramiento es consistente con la proporción detectada; por tanto sólo presentaremos los resultados para la proporción promedio de *outliers* detectados.

# 6. Resultados

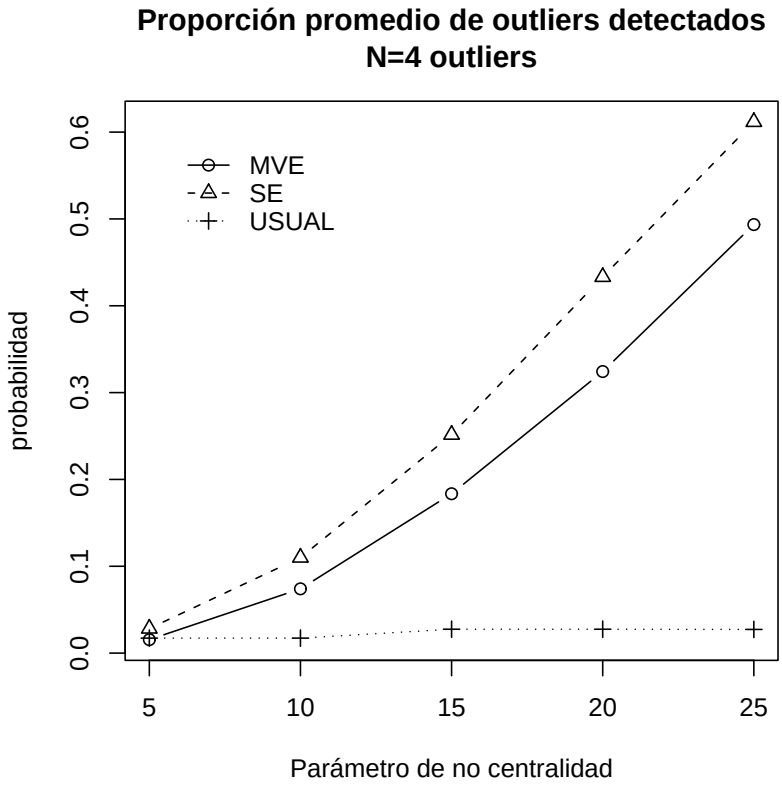
A continuación se ilustran los resultados para la proporción promedio de detección, para $N$ =1, 4 y 7.

## 6.1. Contaminación con desviación de la media

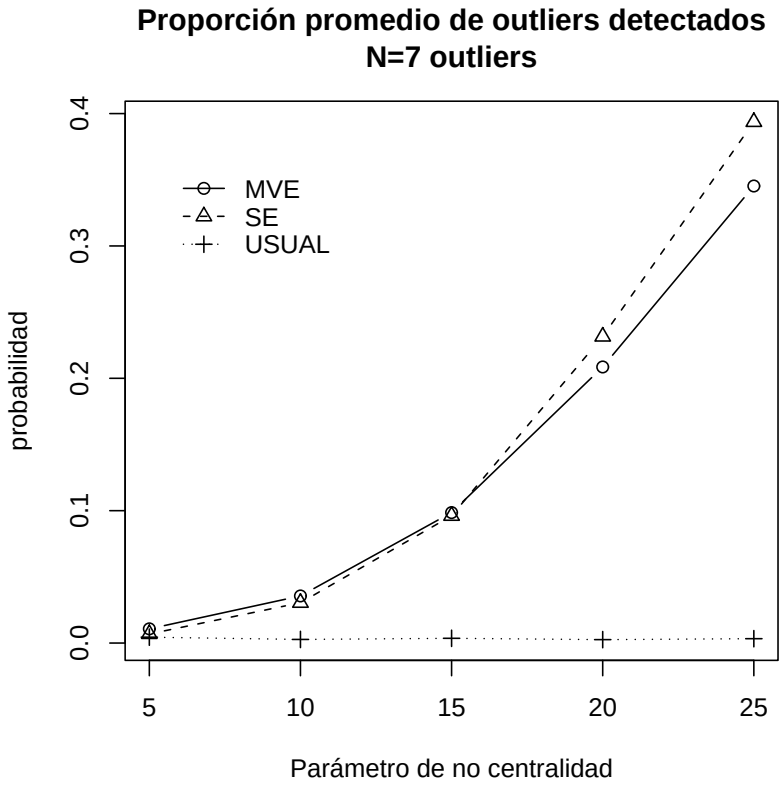
Las figuras 1 a 3 ilustran comparativamente el **pod** de los tres procedimientos. Como es de esperarse, para $N = 1$ *outlier* el USUAL es el mejor de los tres procedimientos. Para $2 \leq N \leq 7$ el SE es el mejor de los tres procedimientos y es claro de la figura 2 que a partir de $N = 4$ el USUAL no detecta en absoluto; también se observa en estas figuras que a medida que $N$ aumenta el procedimiento MVE tiende a aproximarse al SE, de tal forma que para $N = 7$ es poca la diferencia entre los dos procedimientos cuando $d^2 \leq 20$.




Gráfica 1: Efectos de la presencia de $N = 1$ *outlier* sobre el **pod**, en contaminación con desviación de la media.




Gráfica 2: Efectos de la presencia de $N = 4$ *outliers* sobre el **pod**, en contaminación con desviación de la media.




Gráfica 3: Efectos de la presencia de $N = 7$ *outliers* sobre el **pod**, en contaminación con desviación de la media.

## 6.2. Contaminación simétrica

Las figuras 4 a 6 presentan comparativamente el efecto del nivel de contaminación y del factor $\lambda$ sobre el **pod**. Puede verse que las curvas de dicha proporción poseen concavidad hacia abajo a diferencia de las curvas en el caso anterior[3]. Para $\lambda \leq 4$ los tres procedimientos presentan aproximadamente el mismo **pod** en el rango de $1 \leq N \leq 7$; para valores de $\lambda \geq 4.5$, a medida que $N$ aumenta también se incrementan las diferencias de manera notoria. Por ejemplo, para $N = 1$ aunque el USUAL supera al SE y al MVE, los dos primeros son muy similares; para $N \geq 2$ el SE es el procedimiento que detecta con mayor proporción en tanto que el USUAL es el menor de los tres; sin embargo este último no resulta tan insensible a la presencia de *outliers* provenientes de este tipo de contaminación como en el caso de la contaminación con desviación de la media; por ejemplo cuando $N \leq 4$ y $\lambda \leq 8.5$ su **pod** es similar al del MVE, y para $N = 7$ aún sigue detectando.

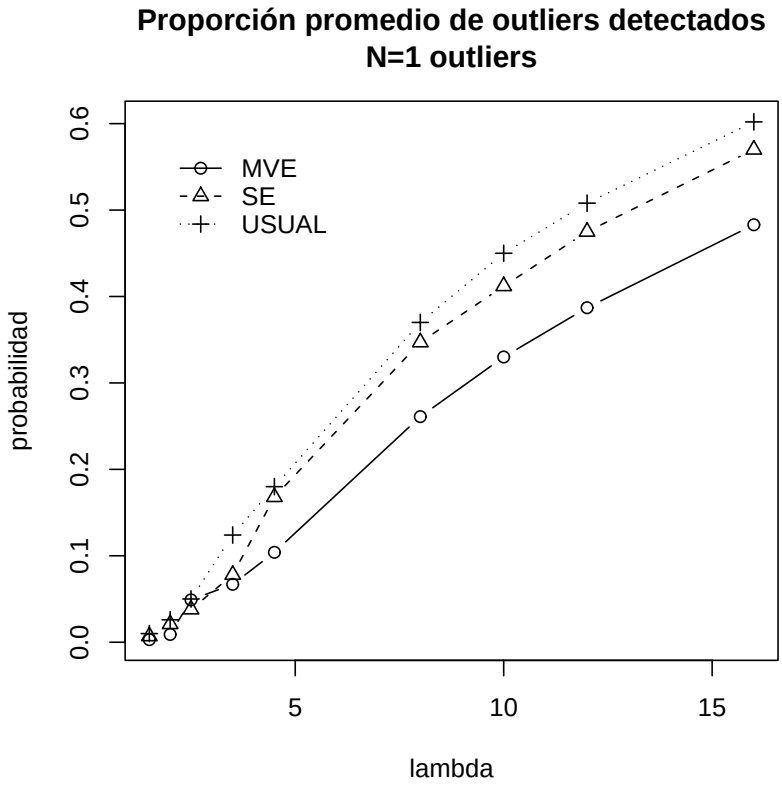
Por su parte, el MVE se mantiene por debajo del SE en todos los niveles de contaminación considerados, y la diferencia entre ambos procedimientos crece con $\lambda$; además, el SE es el único de los tres procedimientos que para $\lambda = 16$ alcanza un **pod** mayor o igual a 0.5 en el rango de contaminación $N \leq 6$.

## 6.3. Contaminación cruzada

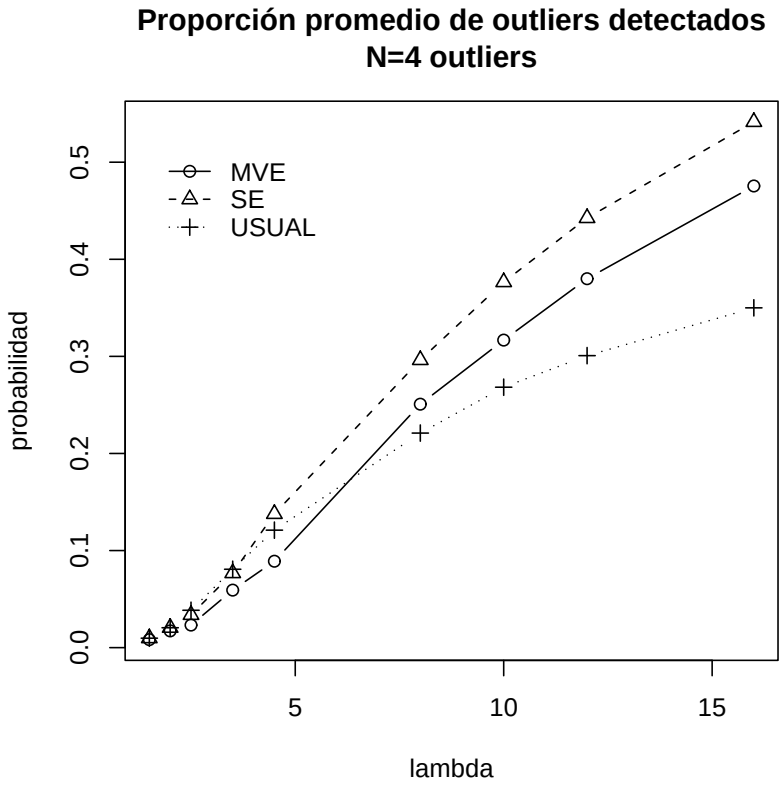
Las figuras 7 a 9 presentan las superficies relativas al **pod** de cada procedimiento. En estas gráficas puede observarse el efecto simultáneo de la variación de los parámetros $d^2$, $\lambda$ y $N$ sobre dicha proporción. Para $\lambda$ fijo, los **pod** *versus* cambios en $d^2$, disminuyen a mayor presencia de *outliers* en la muestra; de la misma forma, con $d^2$ fijo, los **pod** *versus* cambios en el factor de inflación $\lambda$, disminuyen a mayor cantidad de *outliers* presentes. Pero ante la presencia de múltiples *outliers*, los procedimientos son más hábiles en detectar *outliers* con desviaciones en la dirección de $\lambda$ que en la dirección de $d^2$; esto se evidencia por la mayor curvatura en las superficies de respuesta en la primera dirección.

Para $N = 1$, $d^2 = 20, 25$ las curvas correspondientes en la dirección de $\lambda$ son cóncavas hacia arriba, y en particular cuando $d^2 = 25$ y $\lambda = 1.5$ los **pod** son altos (de 0.78 para el USUAL, 0.72 para el SE y 0.63 para el MVE; estos valores obtenidos en las simulaciones no se visualizan directamente en la gráfica). Sin embargo, al aumentar el número de *outliers* en la muestra la característica anterior tiende a desaparecer con mayor rapidez en el caso del procedimiento

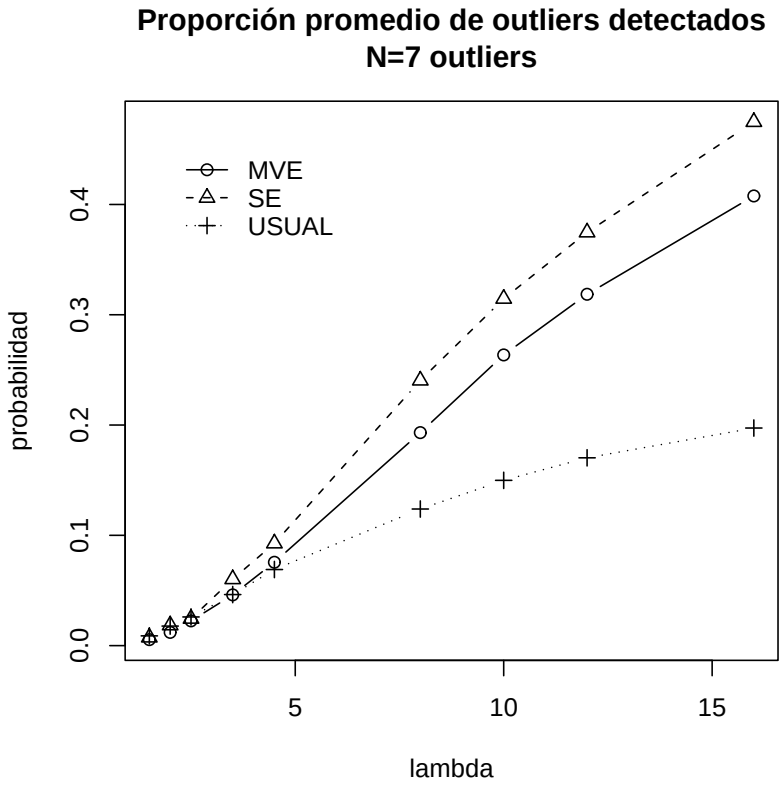
[3]Posiblemente esto se deba a una mayor sensibilidad a cambios en este factor, como se verá mas adelante en el numeral 6.3.




Gráfica 4: Efectos de la presencia de $N = 1$ *outlier* sobre el **pod**, en contaminación simétrica.




Gráfica 5: Efectos de la presencia de $N = 4$ *outliers* sobre el **pod**, en contaminación simétrica.




Gráfica 6: Efectos de la presencia de $N = 7$ *outliers* sobre el **pod**, en contaminación simétrica.

USUAL (cuando $N = 2$)[4] que en otros dos procedimientos (aproximadamente cuando $N = 4$), y al tiempo que disminuye la proporción de detección, la curva de respuesta de SE y MVE para cambios en $d^2$ con $\lambda = 1.5$, va adoptando una ligera concavidad hacia arriba, en tanto que para el USUAL corre casi paralela al eje $d^2$ con proporciones cercanas a cero.

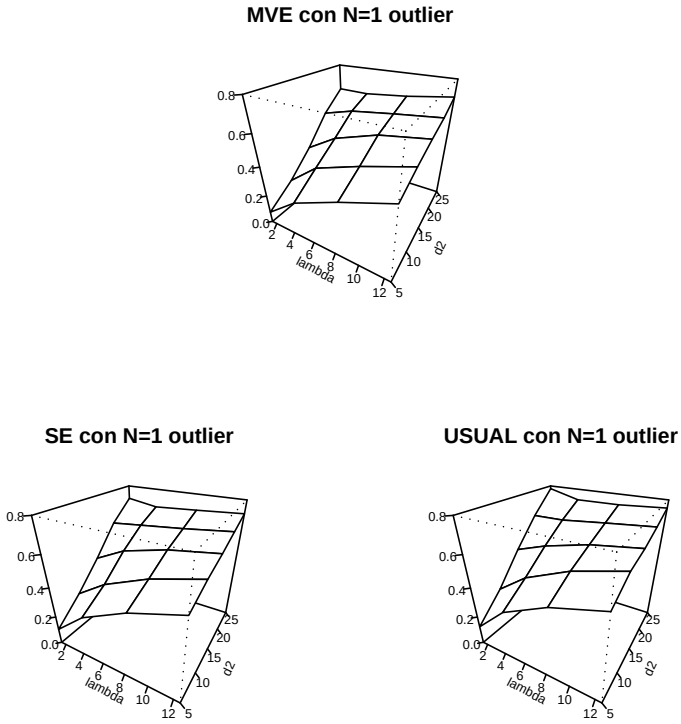
También se observa que en la presencia de múltiples *outliers*, a niveles fijos de $\lambda$, los procedimientos robustos responden aún a cambios en el parámetro $d^2$, en tanto que el USUAL se caracteriza por una respuesta constante, es decir, se vuelve insensible a cambios en dicha dirección.

# 7. Conclusión

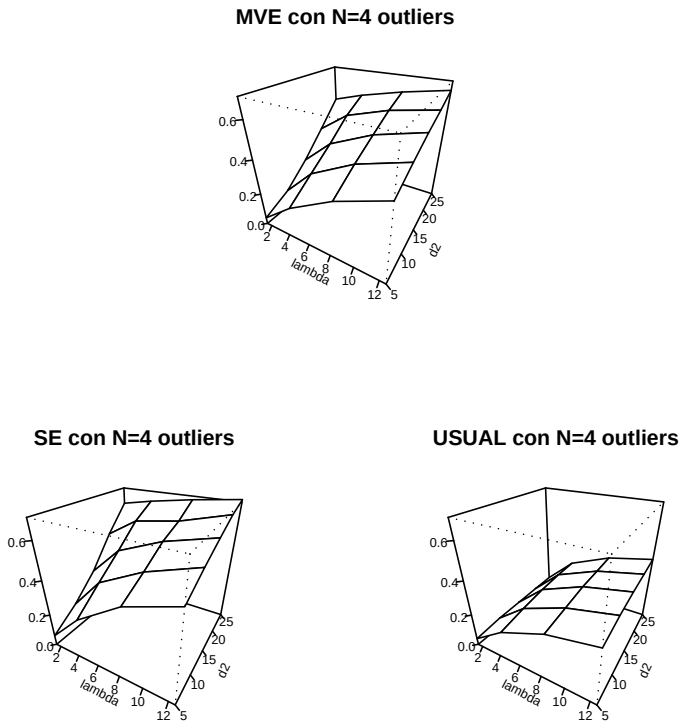
En este artículo se muestra la robustez de los procedimientos SE y MVE en la medida en que son consistentes para la detección de *outliers* originados por los tres tipos de contaminación vistos y ante cambios simultáneos en ambos parámetros $\mu$ y $\mathbf{\Sigma}$. Por tanto, podemos afirmar que en la etapa 1 de la fase I de control con $p = 2$, los estimadores S dentro de los procedimientos considerados constituyen la mejor alternativa para la construcción de una carta $T^2$ robusta ante la presencia de *outliers* ubicados arbitrariamente en el conjunto de datos. Una segunda alternativa, la constituyen los estimadores MVE obtenidos mediante algoritmos genéticos. El procedimiento con estimadores S probó su competitividad con el procedimiento usual cuando hay sólo un *outlier*, y su superioridad cuando hay dos o más *outliers* en el conjunto de datos, tanto de tipo desviación de la media, como en el caso de contaminación simétrica y cruzada.

Se observaron escenarios y condiciones donde el procedimiento con estimadores MVE competía con el S; sin embargo, en la práctica la recomendación general de este trabajo es usar los estimadores S.

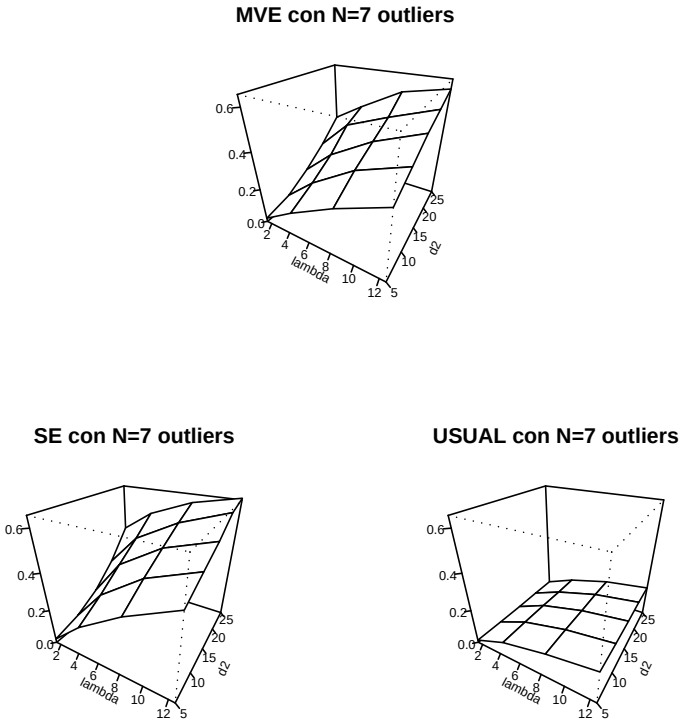
[4]Los resultados gráficos, que no se ilustran aquí, permitieron visualizar esta característica.




Gráfica 7: Efectos de la presencia de $N = 1$ *outlier* sobre el **pod**, en contaminación cruzada.




Gráfica 8: Efectos de la presencia de $N = 4$ *outliers* sobre **pod**, en contaminación cruzada.




Gráfica 9: Efectos de la presencia de $N = 7$ *outliers* sobre el **pod**, en contaminación cruzada.

# Bibliografía

Alt, F. B. & Smith, N. D. (1988), *Multivariate Process Control*, Vol. 7, Handbook of Statistics, pp. 333–351.

Becker, C. & Gather, U. (1999), 'The masking breakdown point of multivariate outlier identification rules', *Journal ofthe American Statistical Association* **94**(447), 947–955.

Gather, U. & Hilker, T. (1997), 'A note on tyler's modificaction of the mad for the stahel-donoho estimator', *The Annals of Statistics* **25**, 2024–2026.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986), *Robust Statistics*, John Wiley & Sons, Inc.

Kosinski, A. S. (1999), 'A procedure for the detection of multivariate outliers', *Computational Statistics & Data Analysis* **29**, 145–161.

Maronna, R. A. & Yohai, V. J. (1995), 'The behavior of the stahel-donoho robust multivariate estimator', *Journal of the American Statistical Association* **90**, 330–341.

Peña, D. (2002), *Análisis de datos multivariables manuscrito*, Universidad Carlos III de Madrid.

Rocke, D. M. & Woodruff, D. L. (1996), 'Identification of outliers in multivariate data', *Journal of the American Statistical Association* **91**, 1047–1061.

Rousseeuw, P. J. & Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc.

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

Sullivan, J. H. & Woodall, W. H. (1996), 'A comparision of multivariate control charts for individual observations', *Journal of Quality Technology* **26**, 398–408.

Vargas, J. A. (2002), Robust estimation in multivariate control charts for individual observations. Sometido a publicación.

Woodruff, D. L. & Rocke, D. M. (1994), 'Computable robust estimation on multivariate location and shape in high dimension using compound estimator', *Journal of the American Statistical Association* **89**, 888–896.