

*Estimación Robusta para Monitorear Perfiles de
Regresión Lineal Múltiple en Fase I*

DIANA RUBRICHE CARDENAS

CÓDIGO: 01832519



UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS
DEPARTAMENTO DE ESTADÍSTICA
BOGOTÁ, D.C.
JUNIO 24 DE 2015

*Estimación Robusta para Monitorear Perfiles de
Regresión Lineal Múltiple en Fase I*

DIANA RUBRICHE CARDENAS

CÓDIGO: 01832519

TESIS DE MAESTRIA PRESENTADA COMO REQUISITO PARCIAL
PARA OPTAR AL TÍTULO DE
MAGISTER EN CIENCIAS ESTADISTICA

DIRECTOR
JOSÉ ALBERTO VARGAS, PH.D.



UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS
DEPARTAMENTO DE ESTADÍSTICA
BOGOTÁ, D.C.
JUNIO 24 DE 2015

Título en español

Estimación Robusta para Monitorear Perfiles de Regresión Lineal Múltiple en Fase I

Title in English

Robust Estimation for Phase I Monitoring of Multiple Linear Regression Profiles

Resumen: En muchas aplicaciones industriales, la calidad de un producto o proceso es mejor representada y resumida por una relación, o perfil, entre una variable respuesta y un conjunto de variables explicativas. Para monitorear este tipo de procesos en Fase I, la carta de control multivariada T^2 de Hotelling es una de las más empleadas. Sin embargo, esta carta de control es muy sensible ante la presencia de datos atípicos. En este documento, se propone utilizar cuatro cartas de control T^2 basadas en estimadores de localización y dispersión robustos. Adicionalmente, a través de simulaciones se comparan estas cuatro cartas de control robustas con la carta T^2 usual. Finalmente, el uso de estos métodos se ilustra con un conjunto de datos reales.

Abstract: In most industrial applications, the quality of a product or process is better characterized and summarized by a relation, or profile, between a response variable and a set of explanatory variables. To monitor these processes in Phase I, multivariate Hotelling's T^2 control chart is often used. However, this control chart is very sensitive to the presence of outliers. In this paper, we propose using four T^2 control charts based on robust estimators of location and dispersion. Additionally, through simulations these four robust control charts are compared with the usual T^2 chart. Finally, the use of these methods are illustrated with a real data set.

Palabras clave: Control estadístico de procesos (SPC), Regresión lineal múltiple, Carta de control multivariada T^2 , Estimación robusta, Perfil.

Keywords: Statistical process control (SPC), Multiple linear regression, Multivariate T^2 control charts, Robust estimation, Profile.

Nota de aceptación

Trabajo de tesis

Aprobado

“Mención Meritoria o Laureada”

Jurado

Jurado

Jurado

Director
José Alberto Vargas Navas

Codirector

Bogotá, D.C., Noviembre 19 de 2014

Dedicado a

Al Diseñador y Hacedor de nuestra asombrosa mente e inteligencia: "Digno eres tú, Jehová, nuestro Dios mismo, de recibir la gloria y la honra y el poder, porque tú creaste todas las cosas, y a causa de tu voluntad existieron y fueron creadas" (Revelación 4:11)

Agradecimientos

Agradezco a mis padres y a mis hermanos por su incondicional apoyo. A Vlady, el amor de mi vida, por la confianza, sus consejos y su amor.

Índice general

Índice general	I
Índice de tablas	III
Índice de figuras	IV
Introducción	VI
1. Control Estadístico de Procesos	1
1.1. Cartas de Control	1
1.1.1. Carta de Control Multivariada T^2 de Hotelling	2
2. Estimadores Robustos de Localización y Dispersión	4
2.1. Estimador MVE	6
2.2. Estimador MCD	7
2.3. Estimador RMCD	9
3. Cartas de Control Robustas	11
3.1. Notación	11
3.2. Cartas de control T^2 robustas para monitorear perfiles	12
3.2.1. Estimación de los límites de control	14
4. Estudio de simulación	17
4.1. Probabilidad de una señal con perfiles atípicos aleatorios	18
4.2. Probabilidad de una señal con un cambio sostenido	20
5. Ejemplos	29
5.1. Ejemplos con datos simulados	29

5.1.1. Ejemplo 1	29
5.1.2. Ejemplo 2	30
5.2. Ejemplo con datos reales	30
A. Programación de los algoritmos	39
A.1. Algoritmo utilizado para fijar los límites de control	39
A.2. Algoritmo utilizado para calcular la probabilidad de una señal fuera de control	41
Conclusiones	46
Trabajo futuro	47
Glosario	48
Bibliografía	49

Índice de tablas

3.1. Variables independientes usadas para el análisis	15
3.2. Límites de control estimados mediante simulación de Monte Carlo	16
4.1. Estimación de la probabilidad de una señal bajo cambios en el intercepto en Y	19
5.1. Parámetros estimados asociados al modelo de regresión lineal múltiple del ejemplo 1 y estadística T^2 usando los estimadores Usual, MVE, MCD, RMCD50 y RMCD75	31
5.2. Parámetros estimados asociados al modelo de regresión lineal múltiple del ejemplo 2 y estadística T^2 usando los estimadores Usual, MVE, MCD, RMCD50 y RMCD75	32
5.3. Resultados de la regresión lineal múltiple para 14 muestras de una aplicación de calibración de la NASA	34

Índice de figuras

2.1. Elipsoide de volumen mínimo para un conjunto de observaciones	7
4.1. Probabilidad de una señal fuera de control bajo cambios en el intercepto en Y de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$, cuando hay m=2,4,6 perfiles atípicos aleatorios	22
4.2. Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_1 a $A_1 + \lambda \frac{\sigma}{\sqrt{S_{xx1}}}$, cuando hay m=2,4,6 perfiles atípicos aleatorios	23
4.3. Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_2 a $A_2 + \lambda \frac{\sigma}{\sqrt{S_{xx2}}}$, cuando hay m=2,4,6 perfiles atípicos aleatorios	24
4.4. Probabilidad de una señal fuera de control bajo cambios en el intercepto en Y de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$, cuando hay m=5,10 perfiles atípicos aleatorios	25
4.5. Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_1 a $A_1 + \lambda \frac{\sigma}{\sqrt{S_{xx1}}}$, cuando hay m=5,10 perfiles atípicos aleatorios	25
4.6. Probabilidad de una señal fuera de control bajo cambios en el intercepto en Y de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$, para cambios sostenidos a partir del 50 % y 75 %	26
4.7. Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_1 a $A_1 + \lambda \frac{\sigma}{\sqrt{S_{xx1}}}$, para cambios sostenidos a partir del 50 % y 75 %	26
4.8. Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_2 a $A_2 + \lambda \frac{\sigma}{\sqrt{S_{xx2}}}$, para cambios sostenidos a partir del 50 % y 75 %	27
4.9. Probabilidad de una señal fuera de control bajo cambios en el intercepto en Y de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$, para cambios sostenidos a partir del 50 % y 75 %	27
4.10. Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_1 a $A_1 + \lambda \frac{\sigma}{\sqrt{S_{xx1}}}$, para cambios sostenidos a partir del 50 % y 75 %	28

5.1. Cartas de control T^2 para datos modificados usando (a) Estimador Usual, (b) Estimador MVE, (c) Estimador MCD, (d) Estimador RMCD50, (e) Estimador RMCD75	35
5.2. Cartas de control T^2 para datos modificados usando (a) Estimador Usual, (b) Estimador MVE, (c) Estimador MCD, (d) Estimador RMCD50, (e) Estimador RMCD75	36
5.3. Cartas de control T^2 para datos de la NASA (a) Estimador Usual, (b) Estimador MVE, (c) Estimador MCD, (d) Estimador RMCD50, (e) Estimador RMCD75	37
5.4. Fuerzas, momentos y balanza aerodinámica	38

Introducción

En muchas aplicaciones del control estadístico de procesos (SPC), es frecuente asumir que la calidad de un proceso o producto se puede representar adecuadamente por la distribución univariada de una característica de calidad o por la distribución multivariada de un vector que contiene diferentes características de calidad. Sin embargo, en algunas situaciones prácticas esta calidad es mejor representada y modelada por una relación funcional entre una variable respuesta y una o más variables explicativas. Esta relación funcional se conoce como un perfil. Tal como lo define Woodall (2007), el monitoreo de perfiles consiste en el uso adecuado de cartas de control, cuando la calidad de un producto o proceso está caracterizada por este tipo de relaciones funcionales, con el objetivo de entender y revisar la estabilidad de estas funciones en el tiempo.

Varios enfoques han sido desarrollados para monitorear perfiles lineales simples y múltiples tanto en Fase I como en Fase II. Por ejemplo, Gulliksen & Wilks (1950) propusieron un método de razón de verosimilitudes para probar la igualdad de los parámetros de regresión en diferentes muestras. Mestek et al. (1994) presentaron una carta de control T^2 de Hotelling basada en vectores que contienen los valores de la variable respuesta Y y consideraron la estabilidad de las curvas de calibración lineal en la determinación fotométrica de Fe^{3+} con ácido sulfosalicílico.

Stover & Brill (1998) propusieron una carta T^2 de Hotelling basada en vectores que contienen los estimadores mínimos cuadrados del intercepto en Y y la pendiente. Kang & Albin (2000) presentaron una carta T^2 de Hotelling muy similar al enfoque de Stover & Brill (1998). Kim et al. (2003) propusieron codificar los valores X dentro de cada muestra de manera que el promedio codificado es cero y aplicar cualquiera de las cartas de control de promedios móviles ponderados exponencialmente (EWMA) en los parámetros de regresión en un conjunto de datos en Fase II, o cartas de control tipo Shewhart en estos parámetros en un conjunto de datos en Fase I. Gupta et al. (2006) compararon el desempeño del enfoque en Fase II usado en el análisis de perfiles lineales propuesto por Croarkin & Varner (1982) con el de Kim et al. (2003).

Mahmoud & Woodall (2004) propusieron usar variables indicadoras en un modelo de regresión lineal múltiple y usar la prueba F para probar la igualdad de los interceptos y las pendientes de las k líneas de regresión. Ellos propusieron usar esta prueba F junto con una carta de control para detectar cambios en la varianza del proceso. Zou et al. (2006a) presentaron una carta EWMA multivariada para monitorear los parámetros de un modelo de perfiles de regresión lineal múltiple en Fase II. Zou et al. (2006b) propusieron una carta de control basada en un modelo de punto de cambio para monitorear este tipo de perfiles. Mahmoud et al. (2007) propusieron un enfoque de punto de cambio basado en técnicas de

regresión segmentadas para probar si los parámetros de regresión se mantienen constantes en un conjunto de datos para perfiles de regresión lineal simple.

Este documento se enfoca en el análisis de perfiles en Fase I, que pueden ser representados adecuadamente por un modelo de regresión lineal múltiple, en el que un conjunto de variables explicativas son usadas para describir el comportamiento de la variable respuesta. En la Fase I del monitoreo de perfiles, uno de los principales análisis a realizar es la identificación de perfiles atípicos. Si las causas asignables de estos perfiles pueden ser determinadas, las muestras son removidas del conjunto de datos, los parámetros son estimados con base en las muestras restantes, y un nuevo límite de control superior es recalculado. Sin embargo, muchos de los métodos que utilizan la carta de control multivariada T^2 basada en los estimadores de localización y dispersión usuales, carecen de robustez ante la presencia de datos atípicos, por un fenómeno conocido como enmascaramiento, el cual inhibe el procedimiento para detectar cualquier señal.

El propósito de este trabajo es presentar métodos robustos ante la presencia de datos atípicos para monitorear perfiles de regresión lineal múltiple. Se propone usar cartas de control T^2 basadas en estimadores de localización y dispersión robustos y eficientes. Consideramos los estimadores del elipsoide de volumen mínimo (MVE) de Rousseeuw & Van Zomeren (1990), propuestos inicialmente por Rousseeuw (1984); los estimadores del determinante de covarianza mínima (MCD) propuesto por Rousseeuw & Driessen (1999), y por Rousseeuw (1984); y los estimadores del determinante de covarianza mínima reponderado (RMCD) propuestos por Rousseeuw & Van Zomeren (1990) y por Woodruff & Rocke (1994). Estos métodos son comparados con la carta de control T^2 estándar basada en los estimadores usuales a través de técnicas de simulación.

Este documento está organizado de la siguiente manera: En el capítulo 1 se presenta el papel que juega el control estadístico de calidad y las herramientas empleadas para el análisis de procesos. El capítulo 2 presenta los métodos de estimación robusta propuestos en este trabajo. En el capítulo 3 se construyen las cartas de control a usar en Fase I basadas en los estimadores robustos. Los resultados de un estudio de simulación para comparar el desempeño de estos esquemas de control en términos de la probabilidad de una señal fuera de control son dados en el capítulo 4. El uso de estos enfoques es ilustrado en el capítulo 5 con tres ejemplos; los dos primeros usan datos simulados para el análisis y el último representa una aplicación de calibración en el Centro de Investigación Langley de la NASA. La última sección contiene los comentarios finales y las conclusiones.

Control Estadístico de Procesos

La calidad puede jugar un papel muy importante en el éxito y la prosperidad de muchas organizaciones manufactureras y de servicios. Una compañía que puede satisfacer las necesidades de los clientes a tiempo, a un precio altamente competitivo y con excelente calidad, puede fácilmente dominar a sus competidores. Por lo tanto, es lógico que las organizaciones consideren la calidad como una estrategia de negocios. La Organización Internacional de Normalización o ISO establece una definición completa de calidad en su sistema de gestión de calidad ISO 9001:2008. De acuerdo con esta norma, la calidad se define como “el grado en el que un conjunto de características inherentes cumple con los requisitos”. Sin embargo, Montgomery (2009) define que la calidad es inversamente proporcional a la variabilidad. Esta definición moderna implica que si la variabilidad respecto a una o varias características de calidad es grande, la calidad del producto disminuye. En este sentido, una reducción de la variabilidad existente en el proceso de producción es considerada también como una mejora de la calidad.

Existen diferentes herramientas y métodos para el mejoramiento de la calidad y la reducción de la variabilidad. El control estadístico de procesos (SPC), es uno de los métodos estadísticos usados extensivamente para alcanzar este objetivo. El SPC se refiere a un conjunto de herramientas potentes que ayudan a monitorear y mejorar la calidad de los productos y servicios al lograr la estabilidad en el proceso y la reducción de la variabilidad. Entre estas herramientas, la carta de control es una de las herramientas más destacadas del SPC.

1.1. Cartas de Control

Las cartas de control fueron introducidas por Walter A. Shewart en el año de 1924, y desde entonces se han aplicado a los procesos en diferentes industrias manufactureras y de servicios. Vargas (2006), define una carta de control como una herramienta estadística usada principalmente para el estudio y control de procesos repetitivos. Es una herramienta útil que grafica las mediciones de una o varias características de calidad contra el tiempo o el número de muestras, con el objetivo de distinguir las causas aleatorias, comunes u ocasionales de variación de las causas de variación asignables o especiales.

Las causas comunes de variación representan la variabilidad inherente o natural que existe en un proceso y son el resultado del efecto acumulado de muchas causas pequeñas, básicamente inevitables. Montgomery (2009) se refiere a esta variabilidad natural como “ruido de fondo”. Un proceso que opera únicamente bajo la presencia de causas comunes se encuentra estadísticamente en control. Por otra parte, la variabilidad que surge de otras fuentes de variación tales como los materiales, el personal, las máquinas, el medio ambiente, etc, se conoce como causas de variación asignables o especiales (Montgomery (2009)). De acuerdo a Deming (1982), las causas de variación especiales se producen por “algo especial, que no hace parte del sistema de causas comunes”. Un proceso que opera bajo la presencia de causas asignables se encuentra estadísticamente fuera de control.

En resumen, las cartas de control permiten monitorear, a través del tiempo, la variabilidad en la calidad de un producto o un proceso, con el objetivo de definir de manera más precisa el estado de control estadístico del mismo, y poder tomar medidas correctivas en caso de encontrar que el proceso no se encuentra en control.

1.1.1. Carta de Control Multivariada T^2 de Hotelling

En muchos procesos industriales, es necesario controlar simultáneamente dos o más características de calidad. Para monitorear este tipo de procesos multivariados de manera que se tome en cuenta la correlación entre las características de calidad, se han implementado cartas de control multivariadas, una de las más empleadas es la carta de control multivariada T^2 de Hotelling (Hotelling (1947), Tracy et al. (1992)).

Para el proceso de construcción de una carta de control multivariada, Alt (1984) y Alt & Smith (1988) han definido dos fases. Tal como lo anotan Woodall (2000) y Woodall et al. (2004), el principal objetivo en la Fase I, es analizar un conjunto histórico de datos para entender la variabilidad de un proceso en el tiempo, evaluar su estabilidad y modelar el proceso bajo control. Este último paso se lleva a cabo usualmente estimando los parámetros del modelo. Por otro lado, la Fase II se relaciona con el monitoreo en línea para detectar rápidamente cambios en los valores estimados de los parámetros en control en la Fase I.

Para implementar la carta de control multivariada T^2 de Hotelling en la Fase I con k observaciones, para cada observación individual j , la estadística T^2 está definida por

$$T^2(j) = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{C}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$$

donde $\mathbf{x}_j = (x_{j1}, \dots, x_{jp})'$, $j = 1, \dots, k$, son kp observaciones en Fase I con media muestral $\bar{\mathbf{x}} = k^{-1} \sum_{j=1}^k \mathbf{x}_j$ y matriz de covarianzas $\mathbf{C} = (k-1)^{-1} \sum_{j=1}^k (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'$. Para cada observación individual, se compara $T^2(j)$ con un límite de control que usualmente se obtiene al asumir que las \mathbf{x}_j son variables independientes provenientes de una distribución normal multivariada, $MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$. Bajo estos supuestos de normalidad e independencia, el límite de control para los datos en Fase I está basado en una distribución Beta y para los datos en Fase II está basado en una distribución F (Tracy et al. (1992)). Si un valor de $T^2(j)$ es mayor que el límite de control superior indica que el proceso ha cambiado por alguna causa de variación.

En la Fase I, se compara la eficacia de los métodos en términos de la probabilidad de decidir si el proceso es estable. Esto es la probabilidad de obtener por lo menos una

estadística fuera de los límites de control usando un conjunto de observaciones históricas. En la Fase II, sin embargo, se compara la eficacia de los métodos en términos de la distribución de la longitud de corrida, donde la longitud de corrida está definida como el número de muestras tomadas hasta que de una señal fuera de control.

Estimadores Robustos de Localización y Dispersión

Se asume que hay k observaciones estadísticamente independientes de una distribución normal multivariada. Es decir, las \mathbf{x}_j son independientes e idénticamente distribuidas (iid) $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ para $j = 1, \dots, k$. Los estimadores usuales de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son el vector de medias muestral

$$\bar{\mathbf{x}} = \frac{1}{k} \sum_{j=1}^k \mathbf{x}_j,$$

y la matriz de covarianzas muestral

$$\mathbf{S} = \frac{1}{k-1} \sum_{j=1}^k (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})',$$

respectivamente. Para observaciones multivariadas individuales, estos estimadores, sin embargo, tienen inconvenientes ante la presencia de datos atípicos. Un solo dato atípico a menudo puede ser identificado por su gran distancia de Mahalanobis, que es una medida que toma en cuenta el vector de medias y la matriz de covarianzas muestral. Sin embargo, la presencia de múltiples datos atípicos puede afectar las estimaciones, especialmente la matriz de covarianzas muestral, hasta el punto que ninguno de los datos atípicos se destaque por tener una gran distancia de Mahalanobis. Así, la presencia de múltiples datos atípicos puede pasar inadvertida en la carta T^2 convencional debido a su efecto en los estimadores, lo cual se conoce como el efecto de enmascaramiento. Se han presentado varias propuestas para tratar este problema, enfocadas en utilizar estimadores que son robustos ante la presencia de múltiples datos atípicos, especialmente estimadores robustos de la matriz de covarianzas. Diferentes autores han publicado sus conclusiones sobre estimación robusta en ajustes multivariados. Por ejemplo, Campbell (1980), Huber (1981), Jensen et al. (2007), Maronna (1976), Rousseeuw (1984), Rousseeuw & Leroy (1987), Vargas (2003) y Variyath & Vattathoor (2013).

Tal como lo define Variyath & Vattathoor (2013), un estimador robusto se considera un buen estimador si este tiene la propiedad de ser afín equivariante junto con un alto punto de ruptura, además de ser muy eficiente.

Más formalmente, los estimadores \mathbf{t}_k y \mathbf{C}_k de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, respectivamente, basados en una muestra aleatoria $\mathbf{x}_1, \dots, \mathbf{x}_k$ de una distribución normal p – variante $MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; son llamados afín equivariantes si, para alguna matriz $\mathbf{A}_{p \times p}$ no singular y un vector $\mathbf{b} \in \mathbb{R}^p$,

$$\begin{aligned}\mathbf{t}_k(\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \dots, \mathbf{A}\mathbf{x}_k + \mathbf{b}) &= \mathbf{A}\mathbf{t}_k(\mathbf{x}_1, \dots, \mathbf{x}_k) + \mathbf{b} \\ \mathbf{C}_k(\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \dots, \mathbf{A}\mathbf{x}_k + \mathbf{b}) &= \mathbf{A}\mathbf{C}_k(\mathbf{x}_1, \dots, \mathbf{x}_k)\mathbf{A}'.\end{aligned}$$

La afín equivarianza de los estimadores es muy importante porque hace que el análisis sea independiente de la escala de medida de las variables así como de transformaciones o rotaciones de los datos.

El punto de ruptura de una muestra finita, introducido por Donoho & Huber (1983), es una medida global muy popular de robustes. Intuitivamente, esta es la cantidad más pequeña de contaminación necesaria para afectar un estimador completamente. Un alto punto de ruptura implica un estimador más robusto. Davies (1987) mostró que el punto de ruptura más alto que se puede alcanzar para un estimador de localización y matriz de dispersión afín equivariante en una muestra finita es $\lfloor (k - p + 1)/2 \rfloor / k$. Un estimador se dice que es relativamente más eficiente que cualquier otro estimador si al comparar sus cuadrados medios del error (CME), este posee un menor CME.

Los estimadores clásicos de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, es decir, el vector de medias muestral y la matriz de covarianzas, son afín equivariantes pero su punto de ruptura para una muestra finita es $1/k$, lo que significa que un solo dato atípico puede dañar completamente las estimaciones. Diferentes estimadores robustos multivariados de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ han sido propuestos en la literatura, por ejemplo los estimadores M (Maronna (1976)), los estimadores de Stahel-Donoho (Stahel (1981), Donoho (1982)), los estimadores S (Rousseeuw & Yohai (1984), Davies (1987), Lopuhaä (1989)), los estimadores del elipsoide de volumen mínimo (MVE) y del determinante de covarianza mínima (MCD) (Rousseeuw (1985)) y los estimadores del determinante de covarianza mínima reponderado (RMCD) (Rousseeuw & Van Zomeren (1990), Lopuhaä & Rousseeuw (1991), Willems et al. (2002)).

En particular, Vargas (2003) propuso usar estimadores robustos para el vector de medias y la matriz de covarianzas en Fase I. Consideró los estimadores del elipsoide de volumen mínimo (MVE) de Rousseeuw & Van Zomeren (1990), propuestos inicialmente por Rousseeuw (1984), y el determinante de covarianza mínima (MCD) método de Rousseeuw & Driessen (1999), también propuesto en Rousseeuw (1984). Con el método MVE, se busca el elipsoide más pequeño que contenga por lo menos la mitad de las observaciones. El estimador MVE de localización es el vector de medias (el centro) de dicho elipsoide, y el estimador de dispersión es la matriz de covarianzas de estas observaciones dentro del elipsoide. Las distancias basadas en estos estimadores han demostrado ser muy efectivas en detectar diferentes datos atípicos en una nube de puntos multivariada (Rousseeuw & Van Zomeren (1990)).

Con el método MCD, se busca un subconjunto que contenga la mitad de los datos cuya matriz de covarianzas tenga el determinante más pequeño. El correspondiente centro y la matriz de covarianzas son los estimadores MCD de localización y dispersión. Si se desea conseguir estimadores más robustos y eficientes, una alternativa es utilizar los estimadores del determinante de covarianza mínima reponderado (RMCD). A continuación se expone con más detalle estos tres estimadores.

2.1. Estimador MVE

Los estimadores del elipsoide de volumen mínimo (MVE) fueron propuestos por primera vez por Rousseeuw (1984). El estimador MVE de localización multivariado, \mathbf{t} , corresponde al centro del elipsoide de volumen mínimo (MVE) que cubre al menos el 50% de las observaciones; 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, pag. 258).

Tal como lo presenta Rousseeuw & Van 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:

$$\#\{j; (\mathbf{x}_j - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{x}_j - \mathbf{t}) \leq a^2\} \geq h$$

donde $h = \lfloor \frac{k+p+1}{2} \rfloor$, $\lfloor q \rfloor$ la parte entera de q y a^2 es una constante que puede tomarse igual a $\chi_{0.5,p}^2$ cuando se espera que la mayoría de los datos vengan de una distribución normal.

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. La Figura 2.1, ilustra este concepto.

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 `cov.mve` de R, basada en un algoritmo genético. A continuación se presenta el algoritmo para calcular los estimadores MVE.

1. Repetir el siguiente procedimiento n veces:

Seleccionar una submuestra aleatoria de $(p+1)$ observaciones diferentes, indexadas por $R = \{j_1, \dots, j_{p+1}\}$. Para esta submuestra calcular la media y la matriz de covarianzas:

$$\bar{\mathbf{x}}_R = \frac{1}{p+1} \sum_{j \in R} \mathbf{x}_j \quad y$$

$$\mathbf{S}_R = \frac{1}{p} \sum_{j \in R} (\mathbf{x}_j - \bar{\mathbf{x}}_R)(\mathbf{x}_j - \bar{\mathbf{x}}_R)'$$

El elipsoide correspondiente es inflado o desinflado para que contenga exactamente h observaciones, lo que corresponde a calcular las distancias

$$d_R^2(j) = (\mathbf{x}_j - \bar{\mathbf{x}}_R)' \mathbf{S}_R^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}_R)$$

y calcular $V_R = m_R^{2p} \det(\mathbf{S}_R)$, donde m_R^2 es la h^{th} estadística de orden de la $d_R^2(j)$. Usualmente $h = \lfloor k + p + 1/2 \rfloor$, donde $\lfloor a \rfloor$ representa la parte entera de a .

2. Guarde la R_* para la cual V_R es mínimo a través de todas las n replicaciones. Los estimadores MVE son

$$\bar{\mathbf{x}}_{MVE} = \bar{\mathbf{x}}_{R_*} \quad \text{y}$$

$$\mathbf{S}_{MVE} = c_{k,p}^2 (\chi_{p,0.5}^2)^{-1} m_R^2 \mathbf{S}_{R_*},$$

donde $c_{k,p}^2$ es un factor de corrección para muestras pequeñas, y $\chi_{p,0.5}^2$ es la mediana de la distribución chi cuadrado con p grados de libertad. El número de submuestras n depende del argumento probabilístico

$$1 - (1 - (1 - \varepsilon)^{p+1})^n \geq p_0$$

donde ε se ha tomado como 0.5 y p_0 , la probabilidad de que por lo menos una de las n submuestras contenga p observaciones buenas, este valor está cerca a uno, por ejemplo 0.95 o 0.99. Al realizar un estudio de simulación Rousseeuw & Van Zomeren (1990) encontraron que un factor de corrección razonable para muestras pequeñas está dado por

$$c_{k,p}^2 = \left(1 + \frac{15}{k-p}\right)^2$$

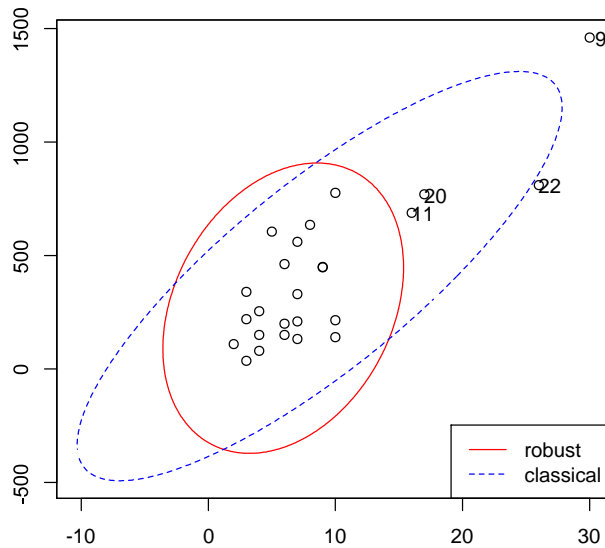


Figura 2.1: Elipsoide de volumen mínimo para un conjunto de observaciones

2.2. Estimador MCD

Sea $\mathbf{x}_1, \dots, \mathbf{x}_k$ una muestra aleatoria tomada de una distribución absolutamente continua F en \mathbb{R}^p . Los estimadores de localización y dispersión MCD, propuestos por primera vez por Rousseeuw (1984), son determinados por el subconjunto de tamaño $h = \lfloor k\gamma \rfloor$ (donde $0.5 \leq \gamma \leq 1$), que da el determinante más pequeño posible de la matriz de covarianzas.

El estimador de localización MCD, $\bar{\mathbf{x}}_{\text{MCD}}$, está definido como el promedio de este subconjunto de h puntos, y el estimador de dispersión MCD está dado por $\mathbf{S}_{\text{MCD}} = a_{\gamma,p}^k \mathbf{C}_{\text{MCD}}$, donde \mathbf{C}_{MCD} es la matriz de covarianzas del subconjunto; la constante $a_{\gamma,p}^k$ es $c_{\gamma,p} \times b_{\gamma,p}^k$ donde $c_{\gamma,p}$ es un factor de consistencia (Croux & Haesbroeck (1999)); y $b_{\gamma,p}^k$ es un factor de corrección de la muestra finita (Pison et al. (2002)). Aquí $1 - \gamma$ representa el punto de ruptura (asintótico) de los estimadores MCD. El estimador MCD alcanza su punto de ruptura más alto posible en una muestra finita cuando $h = \lfloor (k + p + 1)/2 \rfloor$ (Rousseeuw & Leroy (1987)).

El procedimiento para llevar a cabo esta estimación ha sido desarrollado en distintos algoritmos que permiten obtener para conjuntos de datos pequeños, el MCD exacto, mientras que, para conjuntos de datos más grandes, un valor aproximado del estimador MCD, entre estos algoritmos puede destacarse el propuesto por Hawkins & Olive (1999) y Rousseeuw & Van Driessen (1999). Este último es el algoritmo más usado y se conoce como el algoritmo FAST-MCD. Este algoritmo está basado en el paso C, que consiste en intercambiar muchas observaciones al mismo tiempo en cada paso, en lugar de intercambiar un par de puntos situados dentro y fuera de la muestra inicialmente establecida. Este algoritmo ha sido implementado en softwares estadísticos tales como SPLUS, R, SAS y Matlab. A continuación se presenta el algoritmo para calcular los estimadores MCD.

1. Repetir el siguiente procedimiento para $i = 1, \dots, n$ (Rousseeuw & Van Driessen (1999) recomendaron $n=500$).

Seleccionar una muestra aleatoria R_i de $p + 1$ observaciones diferentes.

Para $q = 0$, definir $H_q = R_i$. Calcular el vector de medias $\bar{\mathbf{x}}_q$ y la matriz de covarianzas \mathbf{S}_q de H_q . Calcular las k distancias de Mahalanobis

$$d_{q,j}^2 = (\mathbf{x}_j - \bar{\mathbf{x}}_q)' \mathbf{S}_q^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}_q) \quad j = 1, \dots, k.$$

Identificar un conjunto de h observaciones con las distancias más pequeñas,

$$H_{q+1} = \{\mathbf{x}_j : d_{q,j}^2 \leq (d_{q,j}^2)_{h:k}\}, \quad \text{donde}$$

$$(d_{q,j}^2)_{1:k} \leq (d_{q,j}^2)_{2:k} \leq \dots \leq (d_{q,j}^2)_{k:k}$$

son las distancias ordenadas y h es la parte entera de $k + p + 1/2$.

Repetir para $q = 1, 2, \dots, q^*$ hasta que $\det(\mathbf{S}_{q^*}) = 0$ o $\det(\mathbf{S}_{q^*}) = \det(\mathbf{S}_{q^*-1})$

Definir $d_{d,i} = \det(\mathbf{S}_{q^*})$, $H_i^* = H_{q^*}$, $\bar{\mathbf{x}}_i^* = \bar{\mathbf{x}}_{q^*}$ y $\mathbf{S}_{i^*} = \mathbf{S}_{q^*}$

2. Identificar las i^* más pequeñas con determinante mínimo, tal que

$$i^* = \inf(l : d_{d,l} \leq d_{d,j}, 1 \leq i \leq n).$$

3. Los estimadores MCD son $\bar{\mathbf{x}}_{i^*}^*$ y $\mathbf{S}_{i^*}^*$.

2.3. Estimador RMCD

Si los estimadores robustos multivariados van a ser usados en inferencia estadística, deben ser eficientes bajo la distribución normal multivariada. Hay usualmente una compensación entre eficiencia y robustez, pero si se está interesado en tener ambas propiedades, la mejor propuesta parece ser estimadores ponderados (Rousseeuw & Van Zomeren (1990), Woodruff & Rocke (1994))

Chenouri et al. (2009) consideraron una versión modificada de los estimadores de localización y dispersión MCD. Propusieron una carta de control T^2 de Hotelling para observaciones individuales basada en estimadores más robustos y eficientes del vector de medias y la matriz de covarianzas, conocidos como estimadores del determinante de covarianza mínima reponderado (RMCD) (Rousseeuw & Van Zomeren (1990), Lopuhaä & Rousseeuw (1991), Willems et al. (2002)).

Los estimadores de localización y dispersión RMCD son afín equivariantes con un alto punto de ruptura, una tasa de convergencia $k^{-1/2}$ y alta eficiencia. Además no se dejan influenciar excesivamente por la presencia de datos atípicos en el conjunto histórico de datos en la Fase I. Los estudios de simulaciones desarrollados por Chenouri et al. (2009) mostraron que la carta de control robusta basada en las estimaciones del RMCD tiene mejor desempeño que la carta de control robusta basada en los estimadores MCD usados en Vargas (2003), Hardin & Rocke (2004, 2005) y Jensen et al. (2007) para detección de datos atípicos en la Fase I.

Los estimadores RMCD de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son el vector de medias ponderado,

$$\bar{\mathbf{x}}_{\text{RMCD}} = \left(\sum_{j=1}^k w_j \mathbf{x}_j \right) / \left(\sum_{j=1}^k w_j \right),$$

y la matriz de covarianzas,

$$\mathbf{S}_{\text{RMCD}} = c_{\eta,p} d_{\gamma,\eta}^{\eta,p} \frac{\sum_{j=1}^k w_j (\mathbf{x}_j - \bar{\mathbf{x}}_{\text{RMCD}})(\mathbf{x}_j - \bar{\mathbf{x}}_{\text{RMCD}})'}{\sum_{j=1}^k w_j},$$

donde los pesos están basados en las distancias robustas

$$D(\mathbf{x}_j) = \sqrt{(\mathbf{x}_j - \bar{\mathbf{x}}_{\text{MCD}})' \mathbf{S}_{\text{MCD}}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}_{\text{MCD}})}$$

Las observaciones con una distancia $D(\mathbf{x}_j)$ por abajo del valor límite q_η , donde q_η es el η -ésimo cuantil de la distribución chi cuadrado con p grados de libertad, se les asigna peso 1, mientras que a todas las otras observaciones se les asigna peso 0, es decir,

$$w_j = \begin{cases} 1 & \text{si } D(\mathbf{x}_j) \leq q_\eta \\ 0 & \text{en otro caso.} \end{cases}$$

Se usa el valor $\eta = 0.975$, el cual fue recomendado y usado por Rousseeuw & Van Driessen (1999). Usando $c_{\eta,p} = \eta / P(\chi_{p+2}^2 \leq q_\eta)$ hace \mathbf{S}_{RMCD} consistente bajo la distribu-

ción normal multivariada. El factor $d_{\gamma,\eta}^{\eta,p}$ es una corrección de la muestra finita dado por Pison et al. (2002).

Cartas de Control Robustas

3.1. Notación

Tal como lo presenta Mahmoud (2008), para un conjunto de datos de k perfiles con p variables explicativas X_1, X_2, \dots, X_p y una variable respuesta Y , la j -ésima muestra tiene la siguiente forma

$$\{(x_{1ij}, x_{2ij}, \dots, x_{pij}, y_{ij}), i = 1, 2, \dots, n_j\} \text{ con } n_j > p, j = 1, 2, \dots, k$$

Para cada muestra se asume que el modelo que relaciona las variables independientes X_1, X_2, \dots, X_p con la variable respuesta Y es

$$\mathbf{y}_j = \mathbf{X}_j \mathbf{A}_j + \boldsymbol{\varepsilon}_j, \quad j = 1, 2, \dots, k,$$

donde el vector respuesta y la matriz explicativa en la j -ésima muestra son

$$\mathbf{y}_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \dots \\ y_{n_j j} \end{pmatrix}, \quad \mathbf{X}_j = \begin{pmatrix} 1 & x_{11j} & x_{21j} & \dots & x_{p1j} \\ 1 & x_{12j} & x_{22j} & \dots & x_{p2j} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n_j j} & x_{2n_j j} & \dots & x_{pn_j j} \end{pmatrix},$$

el vector de parámetros y el vector de términos del error son $\mathbf{A}_j = \begin{pmatrix} A_{0j} \\ A_{1j} \\ \dots \\ A_{pj} \end{pmatrix}$, y $\boldsymbol{\varepsilon}_j = \begin{pmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \dots \\ \varepsilon_{n_j j} \end{pmatrix}$,

donde los ε_{ij} son variables aleatorias normales independientes, e idénticamente distribuidas (i.i.d) con media cero y varianza σ_j^2 . Se ha asumido que los valores X son constantes conocidas y toman el mismo conjunto de valores para cada muestra. Si los valores X son aleatorios, entonces bajo ciertas condiciones uno puede usar los métodos para Fase I simplemente al modificar las fórmulas para representar la variación de los valores X de muestra a muestra (Mahmoud et al. (2007)). Si el proceso está bajo control, entonces los parámetros de regresión son constantes con $A_{0j} = A_0, A_{1j} = A_1, \dots, A_{pj} = A_p$, y $\sigma_j^2 = \sigma_0^2, j = 1, 2, \dots, k$. Cuando una causa asignable altera el proceso, por lo menos uno de estos parámetros estará afectado.

Para el análisis en Fase I se considera que los valores de los parámetros en control $A_0, A_1, A_2, \dots, A_p$ y σ^2 son desconocidos. Los estimadores de mínimos cuadrados de los parámetros de regresión para la muestra j están dadas por

$$\hat{\mathbf{A}}_j = \begin{pmatrix} a_{0j} \\ a_{1j} \\ \dots \\ a_{pj} \end{pmatrix} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}_j;$$

ver, por ejemplo, Myers (1990, Ca. 2). Además, σ_j^2 es estimado por el j -ésimo error cuadrático medio MSE_j , donde $MSE_j = SSE_j / (n_j - p - 1)$ y $SSE_j = \sum_{i=1}^{n_j} e_{ij}^2$ es la suma de cuadrados residual, donde

$$e_{ij} = y_{ij} - a_{0j} - a_{1j}x_{1i} - a_{2j}x_{2i} - \dots - a_{pj}x_{pi}, \quad i = 1, 2, \dots, n.$$

Los estimadores de mínimos cuadrados de los parámetros de regresión bajo las condiciones en control tiene distribución normal $(p + 1)$ -variada con el vector de medias

$$\boldsymbol{\mu} = (A_0, A_1, A_2, \dots, A_p)^t$$

y la matriz de varianzas covarianzas

$$\boldsymbol{\Sigma} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

También, la cantidad $(n - p - 1)MSE_j / \sigma^2$ está distribuida como una variable aleatoria chi cuadrado con $(n - p - 1)$ grados de libertad independientemente de los estimadores de los parámetros de regresión.

Una vez se obtenga un conjunto de datos bajo control, se estiman los parámetros del proceso en control usando los promedios de los estimadores, es decir,

$$\bar{a}_r = \sum_{j=1}^k a_{rj} / k, \quad r = 0, 1, 2, \dots, p$$

y la varianza en control está estimada por el error cuadrático medio

$$MSE = \sum_{j=1}^k MSE_j / k.$$

3.2. Cartas de control T^2 robustas para monitorear perfiles

Diversos enfoques para el análisis de perfiles de regresión lineal múltiple en Fase I han sido propuestos. Todos estos enfoques son extensiones de los métodos que existen para monitorear procesos de perfiles lineales simples. Uno de estos métodos es la carta de control T^2 multivariada propuesta por Stover & Brill (1998), que se basa en vectores que contienen los estimadores del intercepto en Y y la pendiente, como dos características de calidad. Las estadísticas T^2 de este método son

$$T_j^2 = (\mathbf{z}_j - \bar{\mathbf{z}})^T \mathbf{S}_1^{-1} (\mathbf{z}_j - \bar{\mathbf{z}}) \quad j = 1, 2, \dots, k,$$

donde $\mathbf{z}_j = (a_{0j}, a_{1j})^T$, $\bar{\mathbf{z}} = (\bar{a}_0, \bar{a}_1)^T$, y $\mathbf{S}_1 = \begin{pmatrix} S_0^2 & S_{01}^2 \\ S_{01}^2 & S_1^2 \end{pmatrix}$. Los a_{0j} y a_{1j} son los estimadores mínimos cuadrados de A_0 y A_1 , y $\bar{a}_0 = \sum_{j=1}^k a_{0j}/k$ y $\bar{a}_1 = \sum_{j=1}^k a_{1j}/k$ son los promedios de los interceptos y las pendientes. Además S_0^2 , S_1^2 y S_{01} son la varianza muestral de los valores a_{0j} , la varianza muestral de los valores a_{1j} y la covarianza muestral entre los valores a_{0j} y a_{1j} , respectivamente. Un LCS apropiado en este caso es

$$LCS = (k-1)^2 B_{1, (k-3)/2, \alpha} / k$$

Extendiendo esta propuesta para el análisis de perfiles lineales múltiples, la estadística T^2 de la carta robusta en Fase I se obtiene a través de un vector que contiene los estimadores de los parámetros $A_0, A_1, A_2, \dots, A_p$ asociados al modelo de regresión y la sustitución de los estimadores usuales insesgados del vector de medias y de la matriz de covarianzas por las estimaciones correspondientes al procedimiento de estimación robusta considerado. La forma general de la estadística T^2 para este método es

$$T_j^2 = (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_0) \quad j = 1, 2, \dots, k,$$

donde $\boldsymbol{\beta}_j = (a_{0j}, a_{1j}, \dots, a_{pj})^t$ es un vector aleatorio que contiene los $p+1$ estimadores de los parámetros de regresión para la muestra j , $\bar{\boldsymbol{\beta}}_0$ es el valor esperado de $\boldsymbol{\beta}_j$ cuando el proceso está en control y $\boldsymbol{\Sigma}_0$ la matriz de covarianzas en control de $\boldsymbol{\beta}_j$.

Teniendo en cuenta lo anterior, para este análisis se proponen las siguientes cinco cartas de control T^2 multivariadas para monitorear perfiles de regresión lineal múltiple en Fase I

- T_{Usual}^2 : Carta T^2 basada en el vector de medias muestral y la matriz de covarianzas muestral.
- T_{MVE}^2 : Carta T^2 basada en los estimadores del Elipsoide de Volumen Mínimo (MVE), estudiada por Vargas (2003) y por Jensen et al. (2007).
- T_{MCD}^2 : Carta T^2 basada en los estimadores del Determinante de Covarianza Mínima (MCD), estudiada por Vargas (2003) y por Jensen et al. (2007).
- T_{RMCD50}^2 : Carta T^2 basada en los estimadores del Determinante de Covarianza Mínima Reponderado (RMCD) con punto de ruptura del 50 %, estudiada por Willems et al. (2002) y por Chenouri et al. (2009).
- T_{RMCD75}^2 : Carta T^2 basada en los estimadores del Determinante de Covarianza Mínima Reponderado (RMCD) con punto de ruptura del 25 %, estudiada por Willems et al. (2002) y por Chenouri et al. (2009).

Las estadísticas T^2 para las cartas de control propuestas son respectivamente

$$T_{Usual,j}^2 = (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}})^t \mathbf{S}_{Usual}^{-1} (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}) \quad j = 1, 2, \dots, k,$$

donde $\bar{\boldsymbol{\beta}} = \frac{1}{k} \sum_{j=1}^k \boldsymbol{\beta}_j$ y $\mathbf{S}_{Usual} = \frac{1}{k} \sum_{j=1}^k (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}) (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}})^t$

$$T_{MVE,j}^2 = (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_{MVE})^t \mathbf{S}_{MVE}^{-1} (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_{MVE}) \quad j = 1, 2, \dots, k,$$

donde $\bar{\boldsymbol{\beta}}_{MVE}$ y \mathbf{S}_{MVE} son las estimaciones de $\bar{\boldsymbol{\beta}}_0$ y $\boldsymbol{\Sigma}_0$, respectivamente, basadas en el método MVE (Rousseeuw & Van Zomeren (1990)).

$$T_{MCD,j}^2 = (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_{MCD})^t \mathbf{S}_{MCD}^{-1} (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_{MCD}) \quad j = 1, 2, \dots, k,$$

donde $\bar{\boldsymbol{\beta}}_{MCD}$ y \mathbf{S}_{MCD} son las estimaciones de $\bar{\boldsymbol{\beta}}_0$ y $\boldsymbol{\Sigma}_0$, respectivamente, basadas en el método MCD (Rousseeuw & Van Zomeren 1990).

$$T_{RMCD50,j}^2 = (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_{RMCD50})^t \mathbf{S}_{RMCD50}^{-1} (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_{RMCD50}) \quad j = 1, 2, \dots, k,$$

donde $\bar{\boldsymbol{\beta}}_{RMCD50}$ y \mathbf{S}_{RMCD50} son las estimaciones de $\bar{\boldsymbol{\beta}}_0$ y $\boldsymbol{\Sigma}_0$, respectivamente, basadas en el método RMCD (Chenouri et al. (2009)).

$$T_{RMCD75,j}^2 = (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_{RMCD75})^t \mathbf{S}_{RMCD75}^{-1} (\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_{RMCD75}) \quad j = 1, 2, \dots, k,$$

donde $\bar{\boldsymbol{\beta}}_{RMCD75}$ y \mathbf{S}_{RMCD75} son las estimaciones de $\bar{\boldsymbol{\beta}}_0$ y $\boldsymbol{\Sigma}_0$, respectivamente, basadas en el método RMCD (Chenouri et al. (2009)).

3.2.1. Estimación de los límites de control

La estadística T_{Usual}^2 , en Fase I, tiene una distribución proporcional a una distribución beta con $(p+1)/2$ y $(k-p-2)/2$ grados de libertad. Por lo tanto, un límite de control superior apropiado es $LCS = \frac{(k-1)^2}{k} B_{(p+1)/2, (k-p-2)/2, \alpha}$ (Stover & Brill (1998)).

En el caso de la distribución de las estadísticas T_{MVE}^2 , T_{MCD}^2 y T_{RMCD}^2 , esta converge en distribución a una distribución χ_p^2 cuando $k \rightarrow \infty$ (Jensen et al. (2007), Serfling (1980) y Chenouri et al. (2009)), pero para muestras pequeñas, el límite de control superior de las cartas debe ser calculado vía simulación.

En esta sección, los LCS en Fase I tanto de la carta usual como para las versiones robustas, son determinados mediante simulación de Monte Carlo fijando una tasa de falsa alarma total de 0.05. La programación del algoritmo se desarrolló en el software R y se usó la función `covMcd` del paquete `rrcov` y la función `cov.mve` del paquete `MASS`. El procedimiento que se siguió para la construcción de los LCS es el siguiente

1. Determinar el número de perfiles o muestras y el número de variables a considerar. En este estudio para $p = 2$ se generaron $k = 20$ perfiles independientes y para $p = 5$ se generaron $k = 50$ perfiles independientes.
2. Generar para cada perfil un modelo en control. Para este caso, se asumió que el modelo en control con los parámetros $A_0 = 0, A_1 = 1, A_2 = 1, A_3 = 1, A_4 = 1$ y

Tabla 3.1: Variables independientes usadas para el análisis

x_1	x_2	x_3	x_4	x_5
0	0.2	0.7	0.3	0
0.2	0.7	0.8	0.5	0.1
0.4	0.8	1	0.6	0.3
0.6	1	1.5	0.9	0.7
0.8	1.5	1.7	1	1.2
1	1.7	2.2	1.4	1.6
1.2	1.8	2.5	1.7	1.9
1.4	1.9	2.6	1.9	2.3
1.6	2	2.7	2	2.4
1.8	2.3	2.8	2.2	2.5

$A_5 = 1$, es

$$y_{ij} = \begin{cases} x_{1i} + x_{2i} + \epsilon_{ij}, & \text{si } p = 2, i = 1, 2, \dots, 10, j = 1, 2, \dots, 20, \\ x_{1i} + x_{2i} + x_{3i} + x_{4i} + x_{5i} + \epsilon_{ij}, & \text{si } p = 5, i = 1, 2, \dots, 10, j = 1, 2, \dots, 50, \end{cases}$$

donde los valores de ϵ_{ij} son variables aleatorias normales iid con media 0 y varianza 1. Los valores fijos usados para las variables independientes se muestran en la Tabla 3.1.

3. Calcular los coeficientes de regresión para cada perfil.
4. Para cada uno de los k perfiles generados, calcular la estadística T^2 , usando el estimador deseado y escoger entre estas estadísticas el valor máximo.
5. Repetir este proceso 5.000 veces.
6. El percentil 95 de estos 5.000 valores máximos es el límite de control superior estimado

Los límites de control obtenidos al seguir el procedimiento descrito se muestran en la Tabla 3.2.

Tabla 3.2: Límites de control estimados mediante simulación de Monte Carlo

$p = 2$					
	Usual	MVE	MCD	RMCD75	RMCD50
k=20	10.453	107.388	84.753	30.183	62.495
$p = 5$					
	Usual	MVE	MCD	RMCD75	RMCD50
k=50	18.795	52.627	115.27	35.257	68.717

Estudio de simulación

El objetivo de la simulación es comparar los métodos propuestos en términos de la probabilidad global de una señal fuera de control, cuando se realizan cambios en uno de los parámetros del modelo e indagar la habilidad de los métodos robustos sobre los métodos usuales para detectar rápidamente estos cambios. En este capítulo se comparan cinco cartas de control para monitorear perfiles de regresión lineal múltiple en Fase I.

Durante la sección citaremos estas cartas con la siguiente notación

- **Usual:** Carta T^2 obtenida mediante estimadores usuales
- **MVE:** Carta T^2 obtenida mediante estimadores MVE
- **MCD:** Carta T^2 obtenida mediante estimadores MCD
- **RMCD50:** Carta T^2 obtenida mediante estimadores RMCD con punto de ruptura del 50 %
- **RMCD75:** Carta T^2 obtenida mediante estimadores RMCD con punto de ruptura del 25 %

En este estudio se consideraron dos escenarios. Para el primero, el número de perfiles generados fue $k = 20$ con $p = 2$ variables independientes. Para el segundo, el número de perfiles fue $k = 50$ con $p = 5$ variables independientes. En cada caso se asumió un modelo en control con $A_0 = 0, A_1 = 1, A_2 = 1, A_3 = 1, A_4 = 1$ y $A_5 = 1$, respectivamente.

$$y_{ij} = \begin{cases} x_{1i} + x_{2i} + \epsilon_{ij}, & \text{si } p = 2, i = 1, 2, \dots, 10, j = 1, 2, \dots, 20, \\ x_{1i} + x_{2i} + x_{3i} + x_{4i} + x_{5i} + \epsilon_{ij}, & \text{si } p = 5, i = 1, 2, \dots, 10, j = 1, 2, \dots, 50, \end{cases}$$

donde los valores de ϵ_{ij} son variables aleatorias normales iid con media 0 y varianza 1. Los valores fijos usados para las variables independientes se muestran en la Tabla 3.1.

Los cambios en los parámetros realizados en este documento son los mismos considerados en Mahmoud (2008). En particular los siguientes cambios en los parámetros fueron considerados:

- De A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$
- De A_1 a $A_1 + \lambda \frac{\sigma}{\sqrt{S_{xx1}}}$
- De A_2 a $A_2 + \lambda \frac{\sigma}{\sqrt{S_{xx2}}}$

Con $\lambda = 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5$, $S_{xxp} = \sum_{i=1}^n (x_{pi} - \bar{x}_p)$ y $\bar{x}_p = \frac{\sum_{i=1}^n x_{pi}}{n}$. Cambios en la desviación estándar no fueron considerados en este ejercicio, debido a que el interés principal se centra en la utilización de la carta T^2 para diferentes estimadores robustos. La variabilidad del proceso se monitorea normalmente mediante un esquema separado. Por ejemplo, una carta R en Fase I y una carta EWMA de residuales en Fase II.

4.1. Probabilidad de una señal con perfiles atípicos aleatorios

Para el caso donde se consideran los perfiles de regresión lineal con dos variables independientes, se generaron $m = 2, 4, 6$ perfiles atípicos de manera aleatoria, considerando cambios por separado en cada uno de los parámetros del modelo. Utilizando los límites de control simulados para $p = 2$ y $k = 20$ que se muestran en la Tabla 3.2, se comparó el desempeño de las cartas de control en términos de la probabilidad de una señal, que es estimada como la proporción de valores T^2 que caen fuera del límite de control superior basada en 100.000 simulaciones.

Los resultados de la simulación cuando se consideran diferentes cambios en los coeficientes del modelo de regresión, A_0, A_1 y A_2 , se presentan en las Figuras 4.1-4.3. Las simulaciones se obtuvieron usando la función `rrcov` del paquete `rmcd` y la función `cov.mve` del paquete `MASS` disponibles en el paquete estadístico R. La programación del algoritmo se muestra en el Apéndice A.

La Figura 4.1 presenta la estimación de la probabilidad global de una señal fuera de control cuando el intercepto en Y cambia de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$. Como se muestra el método RMCD75 tiene un mejor desempeño comparado con los otros métodos, aunque la probabilidad de detección no es alta. El método usual es más eficiente que los métodos MCD, RMCD50 y RMCD75 únicamente para $m = 2$, pero para $m = 4, 6$ es el más deficiente. En general ningún método tiene una alta probabilidad de señal cuando se consideran cambios pequeños en el intercepto en Y . Sin embargo como lo muestra la Tabla 4.1, cuando λ toma los valores 6, 8 y 10 la eficacia de los métodos robustos es alta en comparación con el método Usual. Por ejemplo, para cuatro perfiles atípicos, $p = 2$ y $\lambda = 10$, la probabilidad de una señal es 0.844 con MVE, 0.64 con MCD, 0.827 con RMCD75, 0.648 con RMCD50, y solo 0.134 para el método Usual.

La estimación de la probabilidad global de una señal fuera de control para diferentes cambios en el coeficiente de regresión A_1 se muestran en la Figura 4.2. Para este tipo de cambio, el método RMCD75 presenta el mejor desempeño para pocos perfiles atípicos ($m = 2$) y va perdiendo su eficacia cuando estos aumentan ($m = 4, 6$). Los métodos MVE y RMCD50 se comportan de manera similar para todos los valores de m y son los mejores métodos cuando el número de perfiles atípicos aumenta. El estimador Usual tiene un pobre desempeño para $m = 2$ y sigue decayendo a medida que m aumenta. Entre los métodos robustos el método MCD tiene el desempeño más pobre para $m = 2, 4$. Resultados similares

Tabla 4.1: Estimación de la probabilidad de una señal bajo cambios en el intercepto en Y

$p = 2$						
	λ	Usual	MVE	MCD	RMCD75	RMCD50
$m = 2$	6	0.322	0.311	0.290	0.302	0.607
	8	0.369	0.581	0.469	0.568	0.839
	10	0.373	0.780	0.635	0.776	0.818
$p = 2$						
	λ	Usual	MVE	MCD	RMCD75	RMCD50
$m = 4$	6	0.123	0.314	0.254	0.279	0.468
	8	0.137	0.612	0.439	0.570	0.626
	10	0.134	0.844	0.640	0.827	0.648
$p = 2$						
	λ	Usual	MVE	MCD	RMCD75	RMCD50
$m = 6$	6	0.049	0.245	0.164	0.221	0.197
	8	0.049	0.480	0.316	0.462	0.259
	10	0.050	0.682	0.487	0.703	0.269
$p = 5$						
	λ	Usual	MVE	MCD	RMCD75	RMCD50
$m = 5$	6	0.214	0.730	0.278	0.388	0.895
	8	0.266	0.943	0.539	0.798	0.989
	10	0.285	0.980	0.774	0.975	0.991
$p = 5$						
	λ	Usual	MVE	MCD	RMCD75	RMCD50
$m = 10$	6	0.061	0.403	0.229	0.361	0.644
	8	0.061	0.653	0.491	0.814	0.783
	10	0.061	0.791	0.761	0.983	0.802

se obtuvieron cuando se consideró diferentes cambios en el coeficiente de regresión A_2 tal como lo muestra la Figura 4.3.

Par el caso donde se consideran los perfiles de regresión lineal con cinco variables independientes, se generaron $m = 5, 10$ perfiles atípicos de manera aleatoria y utilizando los límites de control simulados que se muestran en la Tabla 3.2, para $p = 5$ y $k = 50$ se comparó el desempeño de las cartas de control en términos de la probabilidad de una señal basada en 100.000 simulaciones.

La Figura 4.4 muestra la probabilidad de una señal fuera de control cuando el intercepto en Y cambia de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$. Al igual que para $p = 2$, se observa que cuando el número de variables independientes aumenta el método RMCD75 tiene una leve mejoría en la detección de perfiles atípicos comparado con los otros métodos, pero en general la probabilidad de detección es muy baja. Tal como lo muestra la Tabla 4.1, el desempeño de los métodos robustos mejora en comparación con el método Usual, cuando λ toma valores más grandes.

La probabilidad de una señal fuera de control para diferentes cambios en el coeficiente de regresión A_1 se muestran en la Figura 4.5. Para este tipo de cambio, el método RMCD75 presenta el mejor desempeño para pocos perfiles atípicos ($m = 5$) y presenta un buen desempeño para $m = 10$. El método RMCD50 presenta el mejor desempeño para $m = 10$. Los métodos MVE y RMCD50 son mejores que el método MCD. El estimador Usual tiene un pobre desempeño para ambos valores de m . Entre los métodos robustos el método MCD tiene el desempeño más pobre para ambos valores de m . Resultados similares se obtuvieron cuando se consideró diferentes cambios en los coeficiente de regresión A_2, A_3, A_4 y A_5 .

4.2. Probabilidad de una señal con un cambio sostenido

Adicionalmente se considera la situación en la que se escoge uno de los parámetro del modelo de regresión para ser contaminado a partir del perfil $k - m$. En este escenario los últimos m perfiles tendrán la misma contaminación. Este caso se conoce como un cambio de nivel en los parámetros. Para este ejercicio se estimó la probabilidad de una señal fuera de control basada en 100.000 simulaciones, para cambios sostenidos en los parámetros del modelo después de 10 y 15 de 20 perfiles para $p = 2$ y después de 25 y 38 de 50 perfiles para $p = 5$.

La Figura 4.6 presenta la estimación de la probabilidad de una señal fuera de control cuando el intercepto en Y cambia de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$ para un cambio sostenido a partir del 50% y el 75% de los perfiles. Tal como muestra la figura, ningún método tiene un buen desempeño ante estos cambios. Contrario a lo que se muestra en la Figura 4.7 donde se consideran cambios en A_1 . En este escenario los métodos MVE y RMCD50 presentan un buen desempeño para $m = 5$. Por ejemplo, para $\lambda = 5$, la probabilidad de una señal es 0.78 con MVE y 0.69 con RMCD50. La eficiencia del método RMCD75 decae para todos los valores de m , además para contaminaciones muy altas todos los métodos muestran un desempeño muy pobre. Resultados similares se presentan en las Figuras 4.9 y 4.10 para $p = 5$.

En general, el estudio de simulación mostró que para pocos perfiles atípicos el mejor método para detectar una probabilidad de señal fuera de control es RMCD75. Para una cantidad mayor de perfiles atípicos los métodos MVE y RMCD50 presentan el mejor

desempeño. El método Usual presenta un desempeño muy pobre en comparación con los métodos robustos. En el caso de un cambio sostenido en los parámetros del modelo, las cartas de control robustas en general son muy eficientes para detectar cambios en los coeficientes de regresión asociados a cada variable cuando se presentan pocas contaminaciones, excepto por la carta RMCD75 que presenta un desempeño muy pobre. En particular los métodos MVE y RMCD50 presentan el mejor desempeño. Contrario a lo que sucede para altas contaminaciones donde ningún método es eficiente.

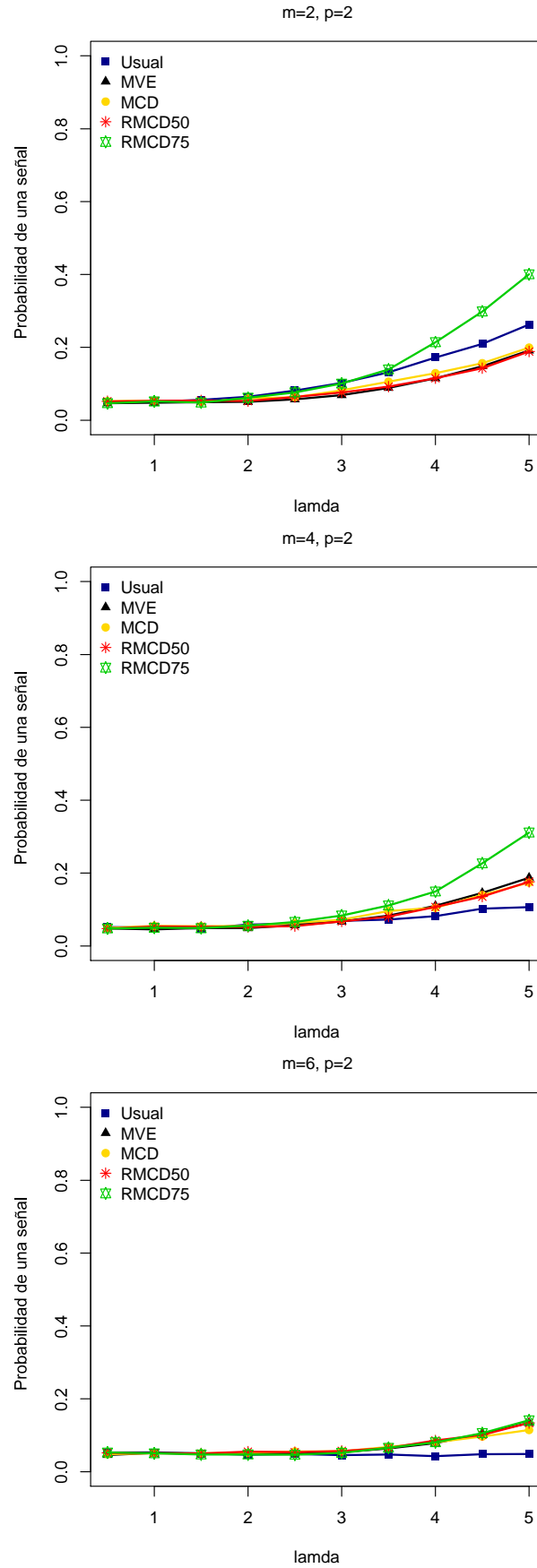


Figura 4.1: Probabilidad de una señal fuera de control bajo cambios en el intercepto en Y de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$, cuando hay $m=2,4,6$ perfiles atípicos aleatorios

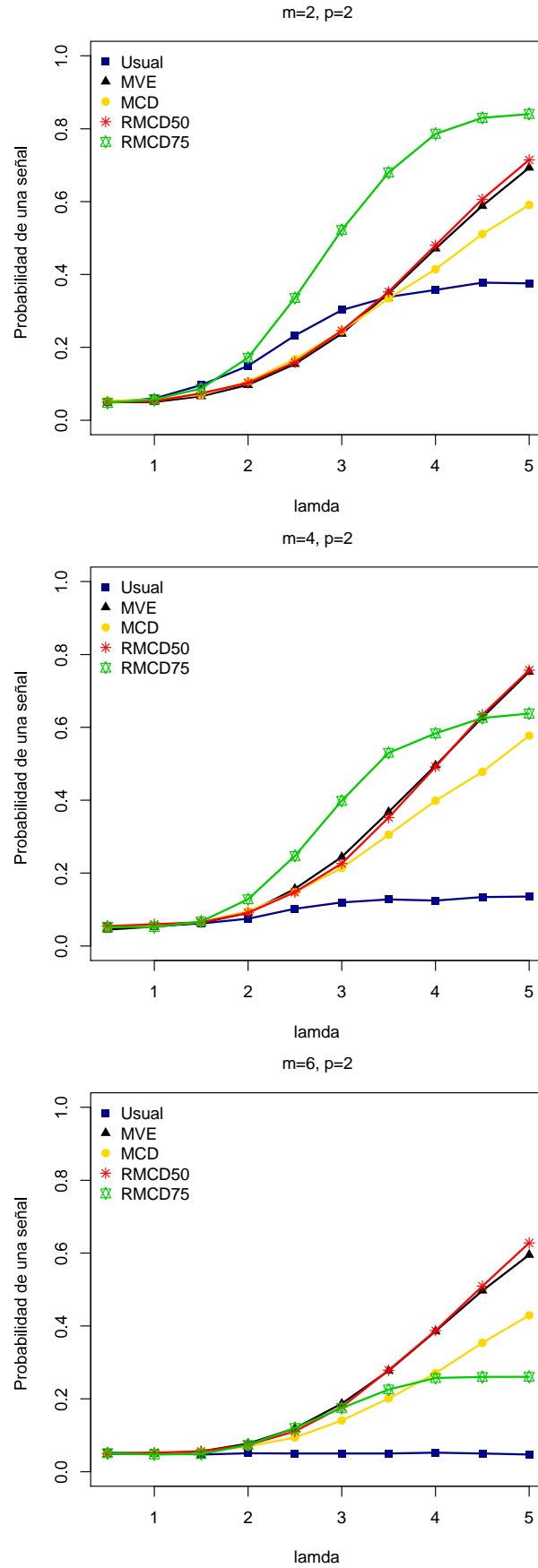


Figura 4.2: Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_1 a $A_1 + \lambda \frac{\sigma}{\sqrt{S_{xx1}}}$, cuando hay $m=2,4,6$ perfiles atípicos aleatorios

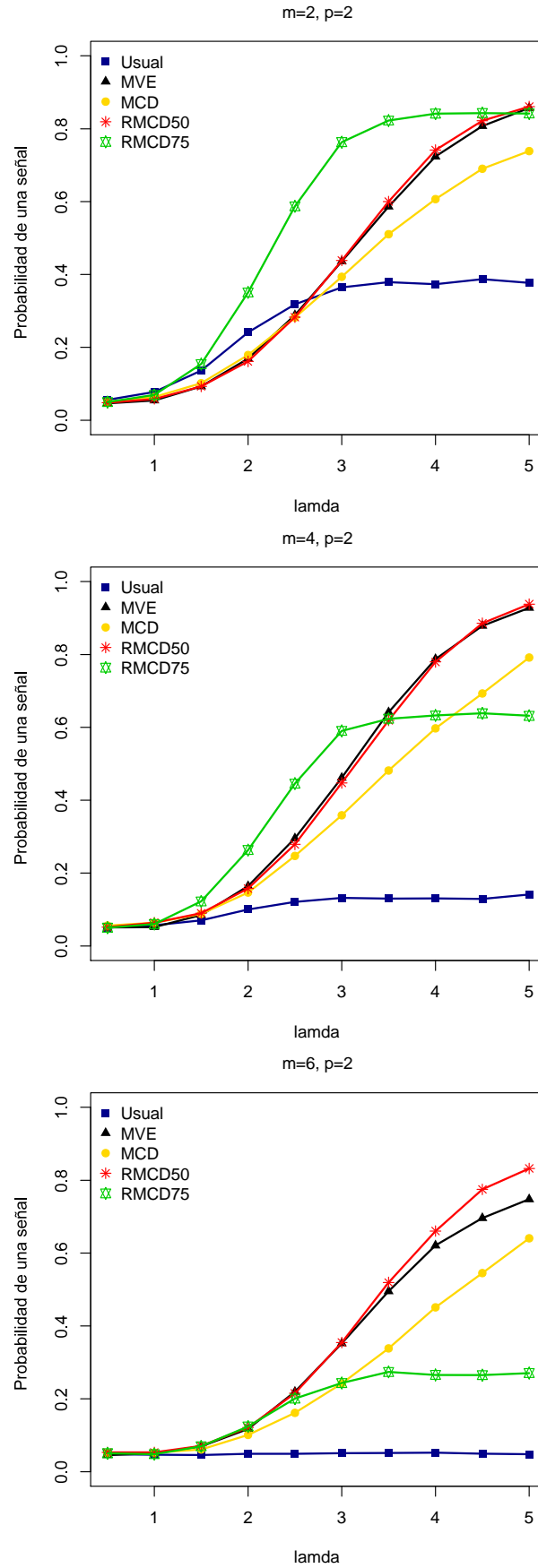


Figura 4.3: Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_2 a $A_2 + \lambda \frac{\sigma}{\sqrt{S_{xx2}}}$, cuando hay $m=2,4,6$ perfiles atípicos aleatorios

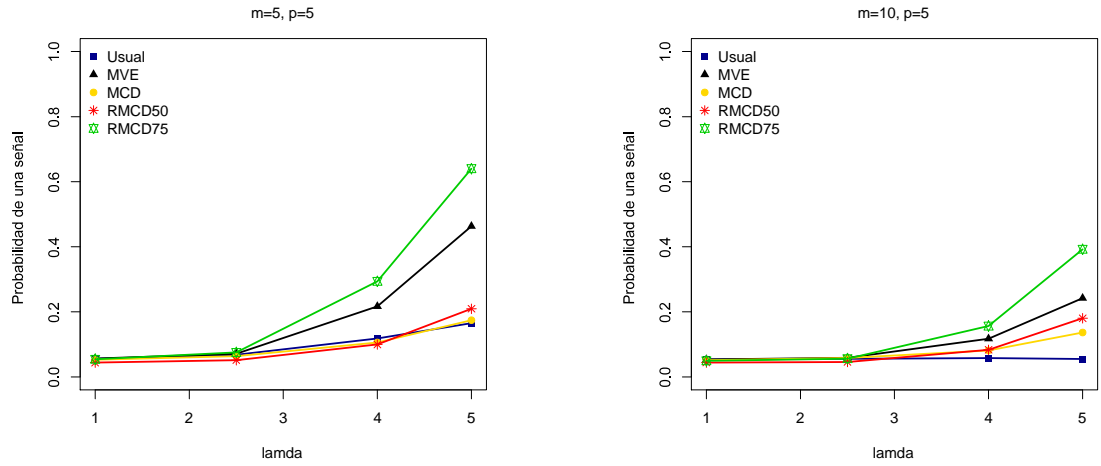


Figura 4.4: Probabilidad de una señal fuera de control bajo cambios en el intercepto en Y de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$, cuando hay $m=5,10$ perfiles atípicos aleatorios

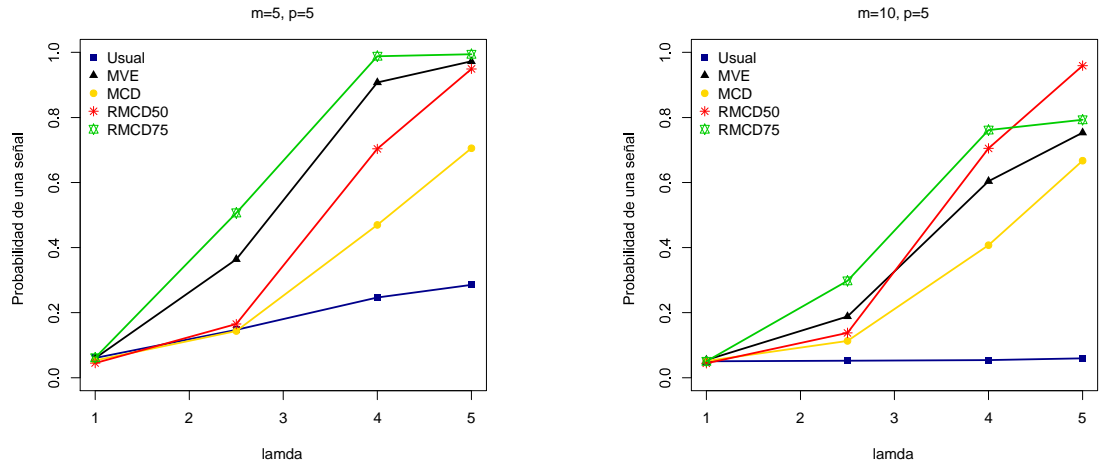


Figura 4.5: Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_1 a $A_1 + \lambda \frac{\sigma}{\sqrt{S_{x,x1}}}$, cuando hay $m=5,10$ perfiles atípicos aleatorios

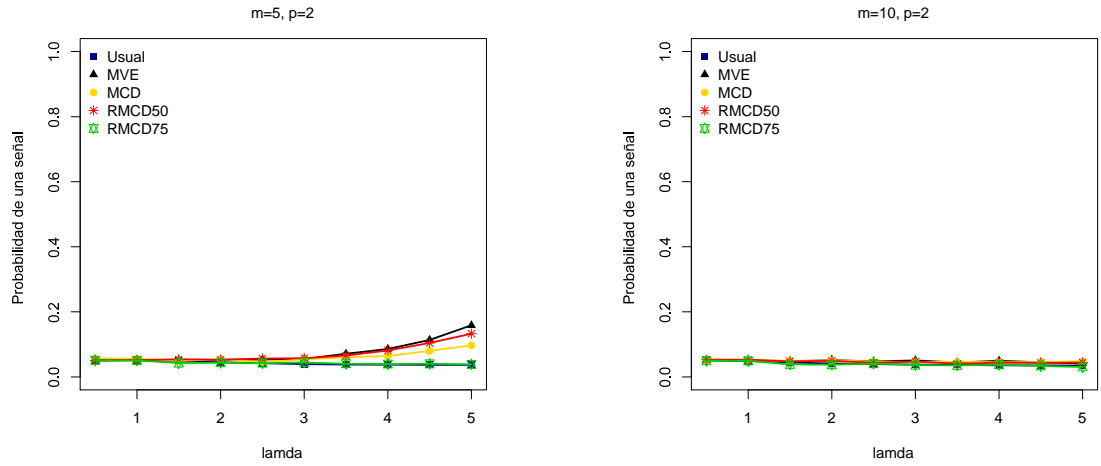


Figura 4.6: Probabilidad de una señal fuera de control bajo cambios en el intercepto en Y de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$, para cambios sostenidos a partir del 50% y 75%

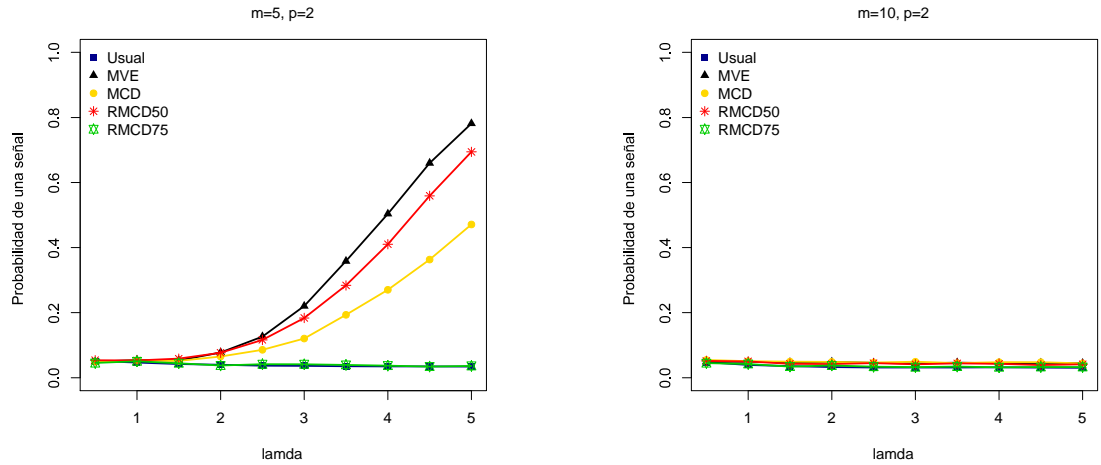


Figura 4.7: Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_1 a $A_1 + \lambda \frac{\sigma}{\sqrt{S_{x \cdot x1}}}$, para cambios sostenidos a partir del 50% y 75%

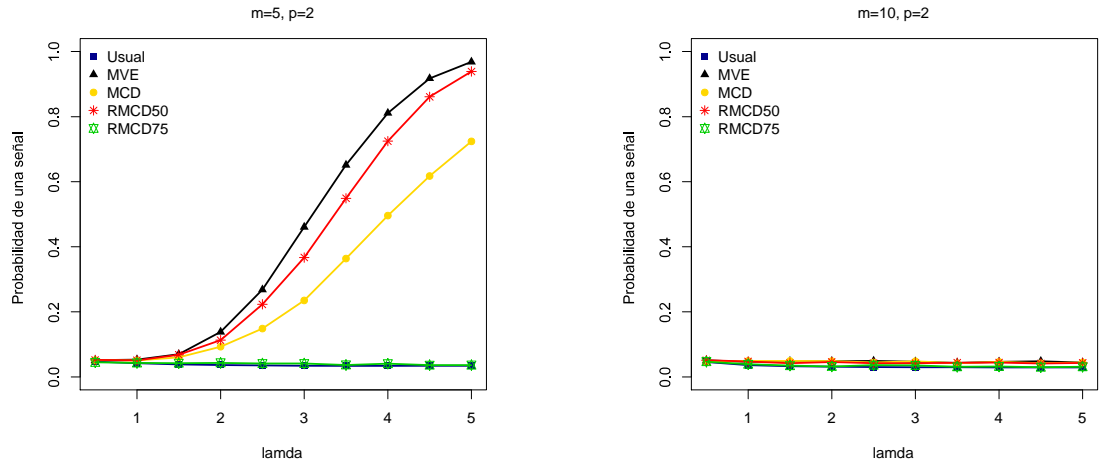


Figura 4.8: Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_2 a $A_2 + \lambda \frac{\sigma}{\sqrt{S_{xx2}}}$, para cambios sostenidos a partir del 50% y 75%

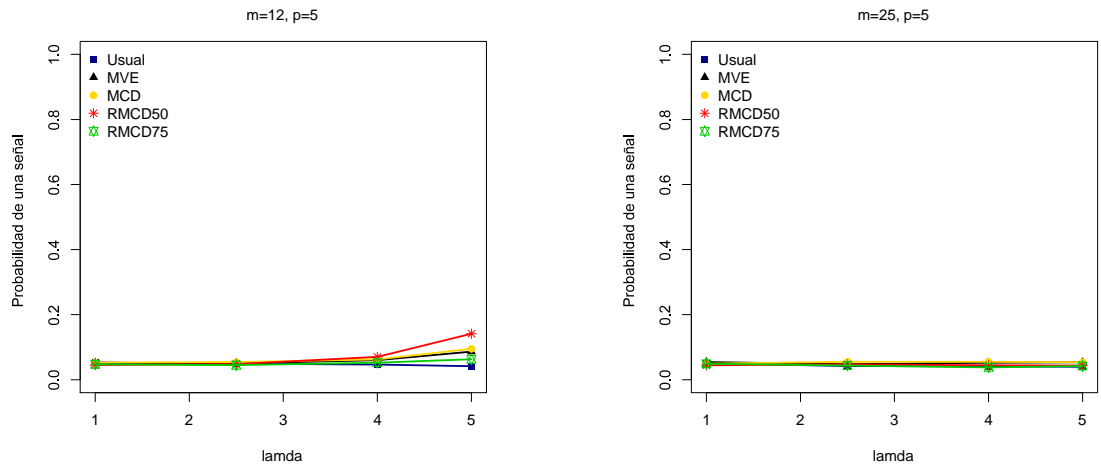


Figura 4.9: Probabilidad de una señal fuera de control bajo cambios en el intercepto en Y de A_0 a $A_0 + \lambda \frac{\sigma}{\sqrt{n}}$, para cambios sostenidos a partir del 50% y 75%

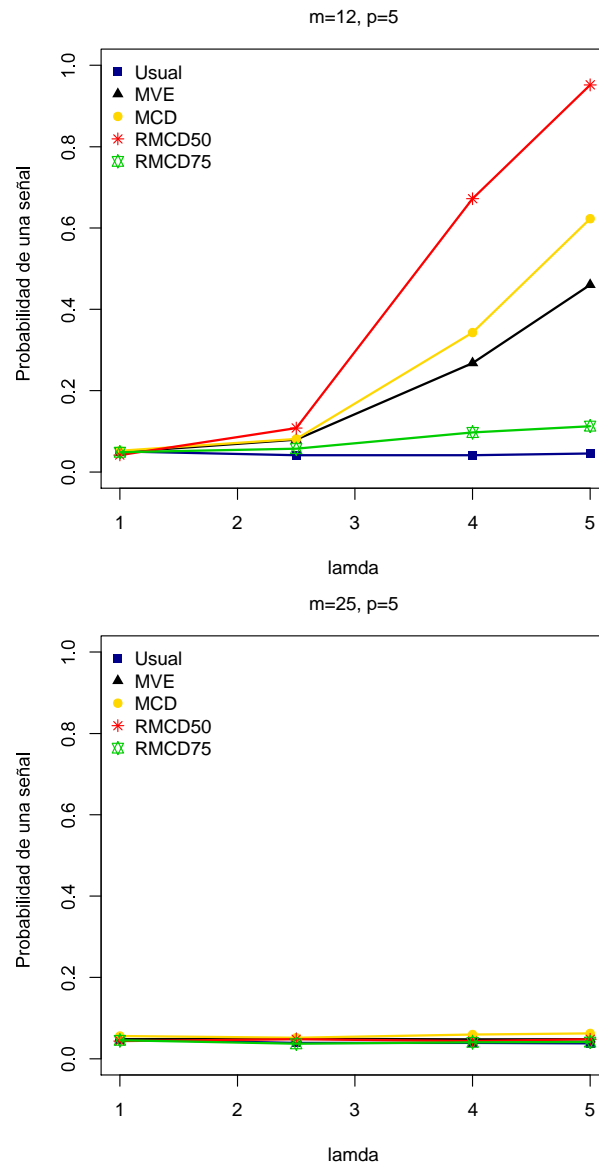


Figura 4.10: Probabilidad de una señal fuera de control bajo cambios en el coeficiente de regresión de A_1 a $A_1 + \lambda \frac{\sigma}{\sqrt{S_{xx1}}}$, para cambios sostenidos a partir del 50% y 75 %

Ejemplos

Este capítulo presenta tres ejemplos. En el primer y segundo ejemplo, dos conjuntos de datos fueron simulados para ser usados en el análisis. En el tercer ejemplo, los métodos robustos introducidos para el análisis de perfiles de regresión lineal múltiple fueron aplicados a un conjunto real de datos.

5.1. Ejemplos con datos simulados

5.1.1. Ejemplo 1

En este ejemplo, 20 muestras de perfiles lineales múltiples fueron generadas de la siguiente manera:

$$y_{ij} = \begin{cases} x_{1i} + x_{2i} + \epsilon_{ij}, & \text{si } i = 1, 2, \dots, 10, \\ & j = 1, 2, \dots, 10, 12, 13, 15, \dots, 20 \\ 3 + x_{1i} + x_{2i} + \epsilon_{ij}, & \text{si } i = 1, 2, \dots, 10, \\ & j = 11, 14, \end{cases}$$

donde los valores de ϵ_{ij} son variables aleatorias normales iid con media 0 y varianza 1. Los valores fijos usados para la variable independiente x_1 fueron 0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6 y 1.8 y para la variable independiente x_2 fueron 0.2, 0.7, 0.8, 1, 1.5, 1.7, 1.8, 1.9, 2 y 2.3. En este ejemplo, fue considerado un cambio de tamaño 3 en el intercepto para los perfiles de regresión para las muestras 11 y 14.

Las estimaciones por mínimos cuadrados para los parámetros de regresión fueron calculadas y aparecen en las columnas 2, 3 y 4 de la Tabla 5.1. Estas estimaciones se usan para comparar los seis métodos estudiados en este documento. Los valores correspondientes de las estadísticas $T_{i,usual}^2$, $T_{i,MVE}^2$, $T_{i,MCD}^2$, $T_{i,RMCD50}^2$ y $T_{i,RMCD75}^2$ basados en estos estimadores aparecen en las columnas 5 a la 9 de la Tabla 5.1.

Comparando estos valores con 10.45, 107.39, 84.75, 62.5 y 30.18, respectivamente, que son los límites de control encontrados mediante simulación para obtener una probabilidad total de falsa alarma de 0.05, se encuentra que todos los métodos robustos señalan los

perfiles 11 y 14 como fuera de control, cuando el tamaño del cambio en el intercepto es grande y se contaminan pocos perfiles. El método usual no es eficiente para detectar este cambio. Las cartas de control usual y T^2 basadas en estos estimadores robustos se muestran en la Figura 5.1.

5.1.2. Ejemplo 2

En el segundo ejemplo, 20 muestras de perfiles lineales múltiples fueron generadas de la siguiente manera:

$$y_{ij} = \begin{cases} 3 + x_{1i} + 2x_{2i} + \epsilon_{ij}, & \text{si } i = 1, 2, \dots, 10 \\ & j = 1, 2, 5, 6, 8, 10, \dots, 14, 16, 18, 19, 20 \\ 3 + x_{1i} + 3.96851x_{2i} + \epsilon_{ij}, & \text{si } i = 1, 2, \dots, 10, \\ & j = 3, 4, 7, 9, 15, 17, \end{cases}$$

donde los valores de ϵ_{ij} son variables aleatorias iid con media 0 y varianza 1. Los valores fijos usados para la variable independiente x_1 fueron 0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6 y 1.8 y para la variable independiente x_2 fueron 0.2, 0.7, 0.8, 1, 1.5, 1.7, 1.8, 1.9, 2 y 2.3. El parámetro asociado a la variable x_2 para seis perfiles, seleccionados de manera aleatoria, fue cambiado de 2 a $2 + \frac{4\sigma}{\sqrt{S_{x_1}}} = 3.96851$.

Tal como en el ejemplo anterior los parámetros estimados de regresión fueron calculadas y aparecen en la columnas 2, 3 y 4 de la Tabla 5.2 y los valores correspondientes de las estadísticas $T_{i,usual}^2$, $T_{i,MVE}^2$, $T_{i,MCD}^2$, $T_{i,RMCD50}^2$ y $T_{i,RMCD75}^2$ basados en estos estimadores aparecen en las columnas 5 a la 9 de la Tabla 5.2.

Al contrastar estos valores con los límites de control simulados, se encuentra que cuando el tamaño del cambio en el parámetro asociado a x_2 es pequeño y el número de contaminaciones es del 30%, únicamente los métodos MVE y RMCD50 señalan muestras fuera de control. En el caso del método MVE detecta todos los perfiles contaminados como fuera de control y en el caso del método RMCD50 señala cinco de seis perfiles como fuera de control, correspondientes a las muestras 4,7,9,15, y 17. El método usual sigue siendo ineficiente para detectar este cambio. Las cartas de control usual y T^2 basadas en estos estimadores robustos se muestran en la Figura 5.2.

5.2. Ejemplo con datos reales

En esta sección, los métodos propuestos en este documento para el análisis de perfiles de regresión lineal múltiple en Fase I, se aplican a un conjunto de datos reales expuestos en Mahmoud (2008) y presentados inicialmente por Parker et al. (2001). El propósito del análisis de este conjunto de datos es investigar las calibraciones replicadas de una balanza de fuerza utilizada en los experimentos del túnel de viento de la NASA.

Una balanza de fuerza permite medir simultáneamente tres componentes ortogonales de la fuerza aerodinámica (normal, axial y lateral) y tres componentes ortogonales del torque aerodinámico (alabeo, cabeceo y guiñada), ejercidos sobre un modelo de avión a escala. Seis respuestas eléctricas son producidas por extensómetros que son proporcionales

Tabla 5.1: Parámetros estimados asociados al modelo de regresión lineal múltiple del ejemplo 1 y estadística T^2 usando los estimadores Usual, MVE, MCD, RMCD50 y RMCD75

M	Intercepto	X_1	X_2	$T_{i,usual}^2$	$T_{i,MVE}^2$	$T_{i,MCD}^2$	$T_{i,RMCD50}^2$	$T_{i,RMCD75}^2$
1	1.42	5.52	-3.05	3.54	2.99	0.79	1.78	2.30
2	1.70	2.10	-1.11	6.81	40.51	11.64	13.05	4.22
3	1.13	4.55	-1.88	1.79	2.16	1.46	1.22	1.90
4	-0.62	-3.60	4.43	2.90	3.68	1.05	1.65	1.85
5	1.23	5.32	-2.83	3.42	3.22	0.93	1.82	2.51
6	0.68	2.66	-0.12	0.77	3.95	3.09	2.24	3.45
7	-0.28	-2.14	3.29	1.42	2.47	0.66	1.09	1.02
8	-0.42	-1.31	3.03	0.64	1.46	0.57	0.67	1.12
9	0.61	1.93	0.04	0.33	0.69	0.27	0.40	0.29
10	-0.42	-1.94	3.09	1.31	1.18	0.75	0.69	0.90
11	4.38	5.86	-2.87	9.04	348.40	205.32	163.05	137.62
12	-0.32	0.34	1.83	0.63	4.78	1.45	1.12	0.50
13	-1.13	-3.75	4.56	2.77	5.20	4.36	2.96	3.55
14	3.65	3.78	-1.22	7.39	293.95	169.08	136.43	114.35
15	-1.93	-2.03	4.36	5.02	54.48	18.49	16.07	3.90
16	1.02	3.52	-1.46	1.68	2.02	0.52	1.10	1.04
17	-0.90	-3.57	4.86	2.37	2.72	0.90	1.58	2.56
18	-0.26	1.82	0.85	2.14	14.23	4.18	3.59	1.36
19	-0.51	-0.92	2.51	0.41	2.48	1.45	0.90	0.51
20	-1.04	0.43	2.05	2.64	34.73	12.06	10.18	2.69

Tabla 5.2: Parámetros estimados asociados al modelo de regresión lineal múltiple del ejemplo 2 y estadística T^2 usando los estimadores Usual, MVE, MCD, RMCD50 y RMCD75

M	Intercepto	X_1	X_2	$T_{i,usual}^2$	$T_{i,MVE}^2$	$T_{i,MCD}^2$	$T_{i,RMCD50}^2$	$T_{i,RMCD75}^2$
1	1.95	0.58	2.96	3.35	3.74	1.09	1.71	2.45
2	2.92	2.06	1.28	1.09	0.84	0.82	0.60	1.41
3	2.82	-2.56	6.29	5.01	140.50	38.87	59.96	4.12
4	3.59	3.13	2.30	3.46	189.08	59.75	84.39	3.03
5	2.51	-0.94	3.62	1.11	0.64	0.30	0.30	0.57
6	3.21	1.12	2.06	0.20	3.49	1.41	1.59	0.30
7	1.56	-0.15	5.92	6.73	206.11	59.04	90.43	6.97
8	4.22	6.99	-2.75	6.66	5.68	10.55	5.30	11.58
9	1.87	-1.04	6.24	3.98	191.31	53.98	82.99	4.42
10	2.23	-1.21	4.06	1.36	1.15	0.55	0.50	0.84
11	3.68	-0.87	2.92	4.95	6.53	1.91	2.98	3.47
12	2.02	-0.73	4.11	1.55	5.32	1.53	2.26	1.18
13	3.07	0.16	2.43	0.90	0.76	0.19	0.30	0.53
14	4.77	6.12	-2.60	5.20	4.72	8.67	3.53	9.65
15	3.67	1.03	3.56	3.57	177.03	52.39	77.01	1.93
16	2.46	-0.49	3.16	1.25	2.24	0.71	0.96	0.59
17	4.02	3.10	1.90	3.74	168.52	53.82	74.84	2.94
18	2.97	-0.19	2.61	1.22	2.44	0.68	1.08	0.67
19	3.55	2.69	0.42	0.97	0.78	1.44	0.56	1.95
20	2.94	0.28	2.41	0.71	0.68	0.16	0.26	0.39

a la magnitud y dirección de sus respectivos componentes aerodinámicos (Ver Figura 5.4). Las relaciones entre seis variables respuestas y seis variables explicativas se modelan utilizando regresión lineal múltiple. Tal como en Mahmoud (2008), para este ejemplo la respuesta axial es la única variable respuesta considerada que se relaciona con las fuerzas aplicadas y los momentos.

El conjunto de datos presentados en Mahmoud (2008) consta de 11 muestras, cada una con n observaciones (64, 73, o 74). A cada una de ellas se le ajustó un modelo de regresión lineal múltiple y se comprobó el cumplimiento de los supuestos del modelo. Mahmoud (2008) reportó que no halló desviaciones del supuesto de normalidad para cada muestra y además encontró que todas las curvas de calibración están bajo control.

Para obtener resultados consistentes al utilizar los estimadores de localización y dispersión robustos que se han propuesto y que se encuentran implementados en el paquete estadístico R, se debe garantizar que el número mínimo de perfiles sea igual a dos veces el número de parámetros a estimar; razón por la cual en este ejemplo en particular se analizan $k = 14$ muestras de perfiles de calibración lineal múltiple. Las muestras 12, 13 y 14 se incluyeron como perfiles atípicos. Los estimadores de mínimos cuadrados de los parámetros de regresión se presentan en la Tabla 5.3.

El límite de control superior para cada una de las cartas que se desean comparar se obtuvo mediante simulación. En este caso utilizando 6 variables explicativas fijas y usando un modelo en control se generaron de manera independiente 14 perfiles. A continuación se estimaron los coeficientes de regresión para cada perfil y con estos valores se calculó la estadística T^2 , usando los estimadores de localización y dispersión deseados. Se escogió entre estas estadísticas el valor máximo. El procedimiento se repitió 5.000 veces y el percentil 95 de estos valores máximos se eligió como límite de control superior estimado.

La Figura 5.3 muestra los cinco métodos aplicados a estos datos. Para el caso de las cartas robustas, los métodos MVE, MCD y RMCD50 señalan todos los perfiles atípicos como fuera de control. Al igual que la carta RMCD75, la carta usual solo es capaz de detectar un perfil fuera de control.

Tabla 5.3: Resultados de la regresión lineal múltiple para 14 muestras de una aplicación de calibración de la NASA

Muestra	Intercepto	Normal	Axial	Balanceo	Cabeceo	Guiñada	Lateral
1	0.4800	0.2370	21.0087	-0.0853	0.0255	-0.1213	0.0106
2	0.4624	0.2368	21.0204	-0.0846	0.0271	-0.1201	0.0096
3	0.4114	0.2363	21.023	-0.0849	0.0262	-0.1215	0.01
4	0.3474	0.2358	21.0253	-0.0853	0.027	-0.1203	0.0101
5	0.4159	0.2349	21.0191	-0.0852	0.0253	-0.1187	0.0109
6	0.4493	0.2368	21.1466	-0.0858	0.0261	-0.1238	0.0122
7	0.4956	0.2370	21.1438	-0.0865	0.0221	-0.1196	0.0116
8	0.8199	0.2357	21.0475	-0.0848	0.0257	-0.1187	0.0109
9	0.5972	0.2353	21.0487	-0.0855	0.0252	-0.119	0.0019
10	0.7336	0.236	21.0491	-0.0851	0.0248	-0.1195	0.0081
11	0.5689	0.2354	21.0518	-0.0852	0.0228	-0.1191	0.0116
12	0.3764	0.2013	20.988	-0.1199	-0.0088	-0.1565	-0.025
13	0.5732	0.2113	21.0247	-0.1095	0.0012	-0.143	-0.0221
14	0.4900	0.2500	21.0200	-0.1000	0.0400	-0.1300	0.0200

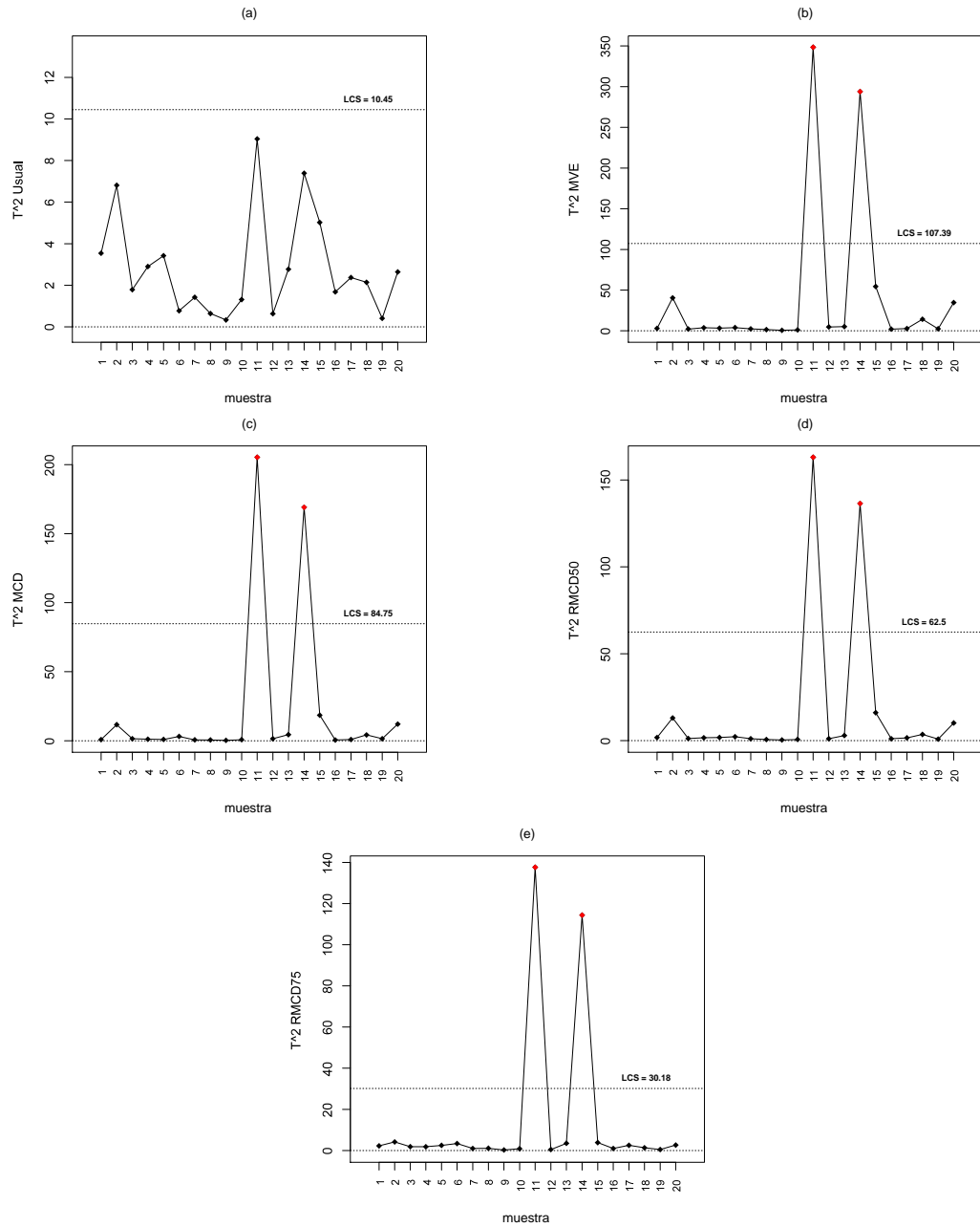


Figura 5.1: Cartas de control T^2 para datos modificados usando (a) Estimador Usual, (b) Estimador MVE, (c) Estimador MCD, (d) Estimador RMCD50, (e) Estimador RMCD75

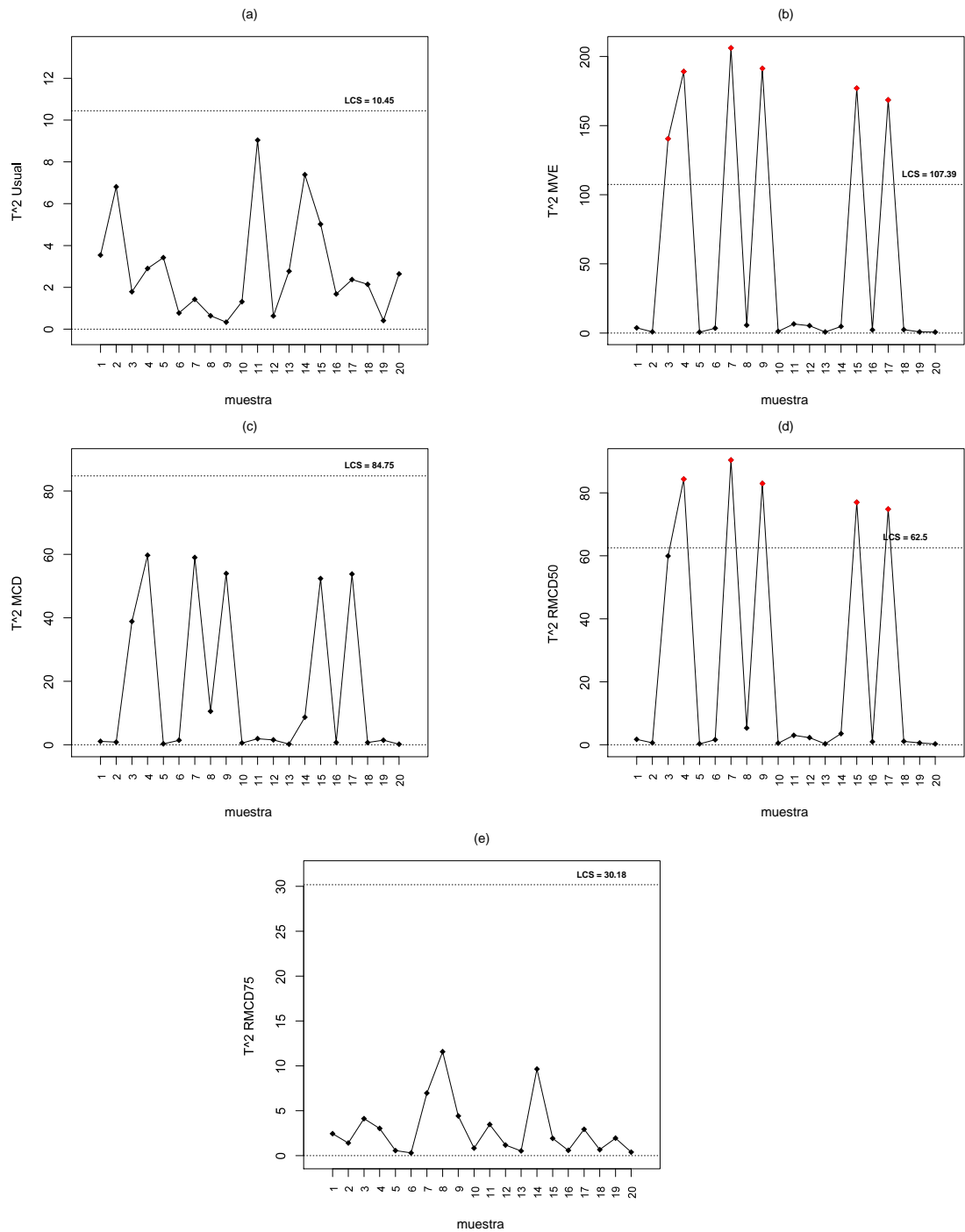


Figura 5.2: Cartas de control T^2 para datos modificados usando (a) Estimador Usual, (b) Estimador MVE, (c) Estimador MCD, (d) Estimador RMCD50, (e) Estimador RMCD75

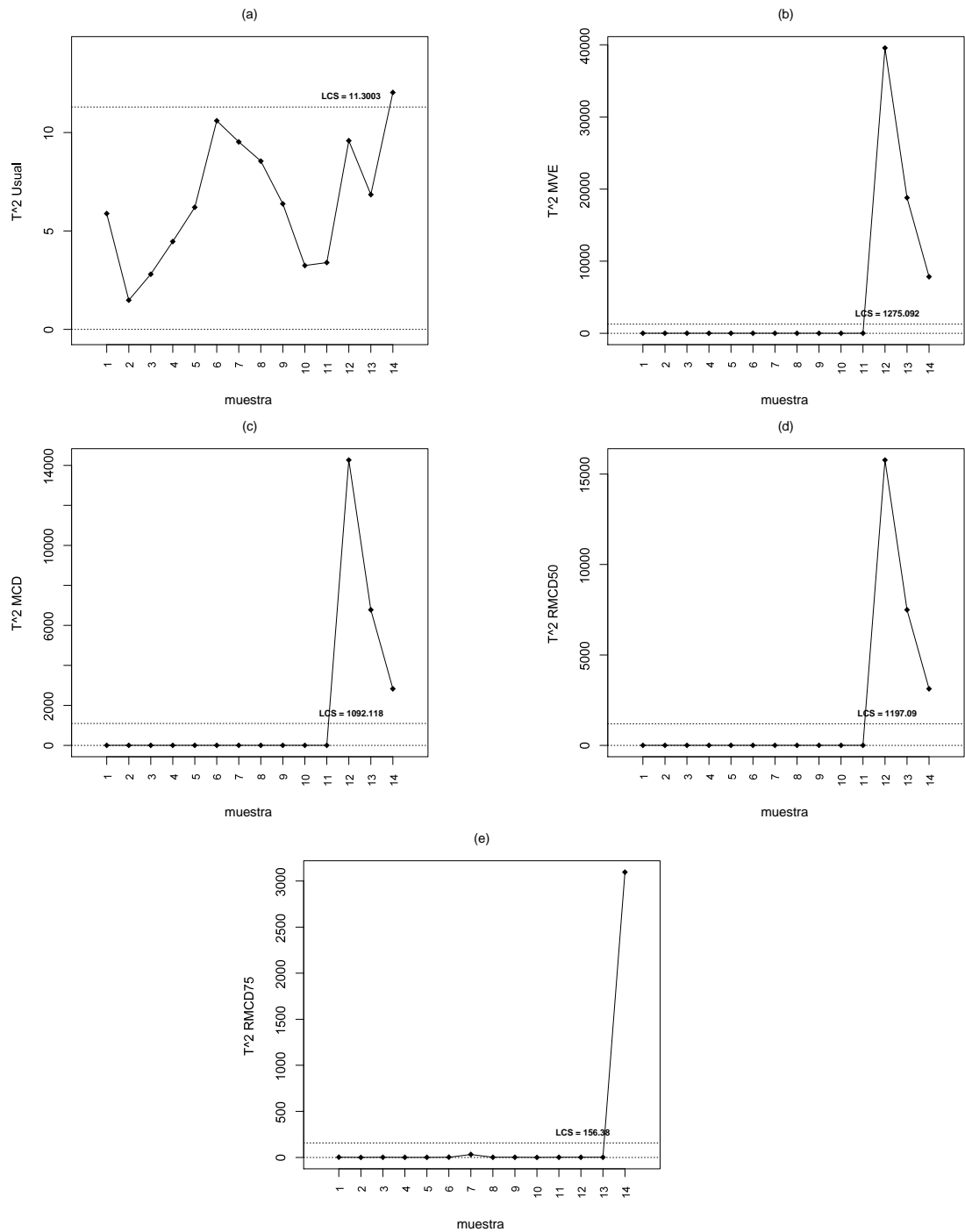


Figura 5.3: Cartas de control T^2 para datos de la NASA (a) Estimador Usual, (b) Estimador MVE, (c) Estimador MCD, (d) Estimador RMCD50, (e) Estimador RMCD75

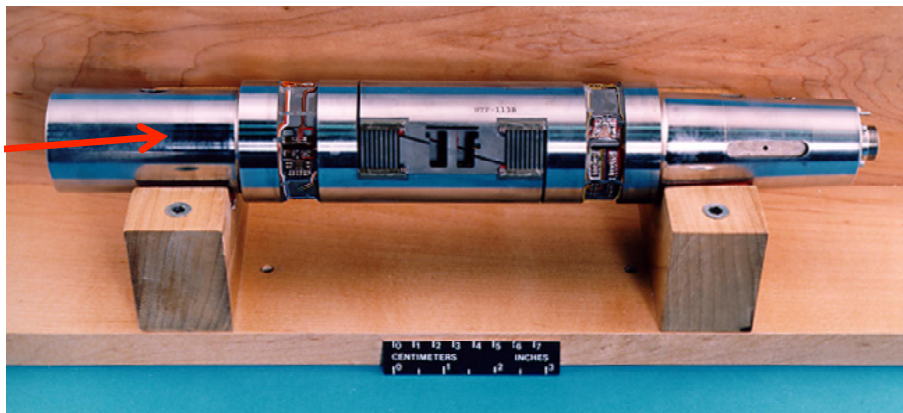
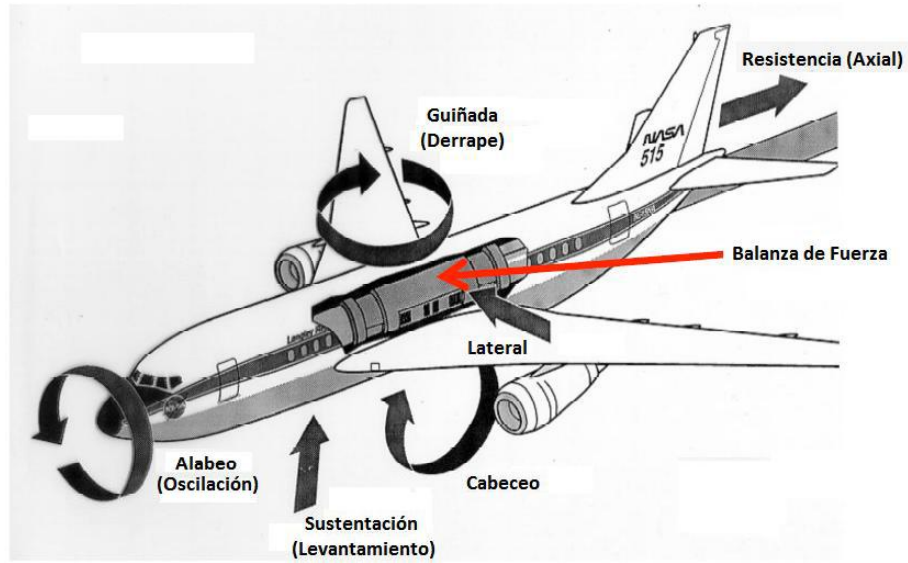


Figura 5.4: Fuerzas, momentos y balanza aerodinámica

Programación de los algoritmos

A.1. Algoritmo utilizado para fijar los límites de control

El siguiente programa calcula el límite de control superior mediante simulación de monte carlo y utiliza la función `CovMcd` del paquete `rrcov` y la función `cov.mve` del paquete `MASS` disponibles en el paquete estadístico R.

```
A0=0 #1
A1=1 #2
A2=1 #3
A3=1 #4
A4=1 #5
A5=1 #6
X1=c(0,0.2,0.4,0.6,0.8,1,1.2,1.4,1.6,1.8) #7
X2=c(0.2,0.7,0.8,1,1.5,1.7,1.8,1.9,2,2.3) #8
X3=c(0.7,0.8,1,1.5,1.7,2.2,2.5,2.6,2.7,2.8) #9
X4=c(0.3,0.5,0.6,0.9,1,1.4,1.7,1.9,2,2.2) #10
X5=c(0,0.1,0.3,0.7,1.2,1.6,1.9,2.3,2.4,2.5) #11
n=length(X1) #12
k= 20 #13
l=5000 #14
p=5 #15

library(MASS) #16
library(rrcov) #17

LCS=function(X1,X2,X3,X4,X5,A0,A1,A2,A3,A4,A5,k,l,p) #18
{
  MX=matrix(0,l,1) #19
  for(g in 1:l) #20
  {
```

```

datos=matrix(0,k,p+1) #21
datos=as.data.frame(datos) #22

for(i in 1:k) #23
{
  Y=A0+A1*X1+A2*X2++A3*X3+A4*X4+A5*X5+rnorm(length(X1),0,1) #24
  reg=lm(Y ~ X1 + X2 + X3 + X4 + X5) #25
  CR=coefficients(reg) #26
  datos[i,]=CR #27
}

s=cov(datos) #28
m=colMeans(datos) #29
#s=cov.mve(datos)$cov #30
#m=cov.mve(datos)$center #31
#s=covMcd(datos)$raw.cov #32
#m=covMcd(datos)$raw.center #33
#s=covMcd(datos,alpha=0.5)$cov #34
#m=covMcd(datos,alpha=0.5)$center #35
#s=covMcd(datos,alpha=0.75)$cov #36
#m=covMcd(datos,alpha=0.75)$center #37

sinv=solve(s) #38

T2=matrix(0,k,1) #39

for(j in 1:k) #40
{
  df=datos[j,]-m #41
  df=as.matrix(df) #42
  T=df%%sinv%%t(df) #43
  T2[j,]=T #44
}

Max=max(T2) #45
MX[g,]=Max #46

}

quantile(MX,0.95) #47
}

LCS(X1,X2,X3,X4,X5,A0,A1,A2,A3,A4,A5,k,1,p)#48
LCS#49

```

Para correr este programa se necesita los siguientes parámetros de entrada

#1-#6: Un intercepto y cinco constantes asociadas a cada variable respectivamente

#7-#11: Cinco variables independientes

#12: n: Tamaño de las variables

#13: k: Número de muestras para cada iteración

#13: l: Número de iteraciones

#15: p: Número de variables

El programa funciona de la siguiente manera

#16: Se carga la librería MASS

#17: Se carga la librería rrcov

#18: Se crea una función llamada LCS que depende de 14 parámetros

#19: Se crea una matriz MX de tamaño $l \times 1$ que guardará el valor de la máxima estadística T2 en cada iteración

#20: Se crea una for para generar 10000 iteraciones

#21: En cada iteración se crea una matriz datos de tamaño $k \times p+1$ que guardara los coeficientes de regresión para cada muestra o perfil

#22: Se crea un data.frame

#23: Se crea una for para generar k muestras en cada iteración

#24: Para cada muestra se genera un modelo Y con los parámetro de entrada

#25-#26: Se estiman los parámetros asociados al modelo

#27: Para cada muestra se almacenan los coeficientes de regresión del perfil en datos

#28-#37: Se elige el estimador deseado y se estima la media y la matriz de varianzas asociadas a las k muestras

#38: Se calcula la inversa de la matriz s

#39: Se crea una matriz T2 de tamaño $k \times 1$ que guardará el valor de la estadística T2 para cada muestra

#40: Se crea una for para generar los valor de la estadística T2 para cada muestra

#41-#43: Se calcula la estadística T2 para cada muestra

#44: Se almacena la estadística T2 para cada muestra en la matriz T2

#45: Se calcula el valor máximo de la estadística T2 para las k muestras

#46: Se almacena el valor máximo en la matriz MX

#47: Se calcula el percentil 95 de todos los valores máximos

#48-#49: Se llama la función LCS

A.2. Algoritmo utilizado para calcular la probabilidad de una señal fuera de control

El siguiente programa calcula la probabilidad de una señal para cambios en los parámetros del modelo, lo hace mediante simulación de monte carlo y utiliza la función CovMcd del paquete rrcov y la función cov.mve del paquete MASS disponibles en el

paquete estadístico R.

```

A0=0 #1
A1=1 #2
A2=1 #3
A3=1 #4
A4=1 #5
A5=1 #6
X1=c(0,0.2,0.4,0.6,0.8,1,1.2,1.4,1.6,1.8) #7
X2=c(0.2,0.7,0.8,1,1.5,1.7,1.8,1.9,2,2.3)#8
X3=c(0.7,0.8,1,1.5,1.7,2.2,2.5,2.6,2.7,2.8)#9
X4=c(0.3,0.5,0.6,0.9,1,1.4,1.7,1.9,2,2.2) #10
X5=c(0,0.1,0.3,0.7,1.2,1.6,1.9,2.3,2.4,2.5) #11
n=length(X1) #12
sx1=var(X1)*(n-1) #13
sx2=var(X2)*(n-1) #14
sig=1 #15
k= 50 #16
l=10000 #17
LCS=35.2576 #18
ml=5 #19
p=5 #20

library(MASS) #21
library(rrcov) #22

SEÑAL=function(X1,X2,X3,X4,X5,A0,A1,A2,A3,A4,A5,k,LCS,ml,sig,lam,p) #23
{
  MX=matrix(0,l,1) #24
  PS=matrix(0,l,1) #25
  for(g in 1:l) #26
  {
    datos=matrix(0,k,p+1) #27
    datos=as.data.frame(datos) #28

    for(i in 1:k) #29
    {
      a=(A0+(lam*sig/sqrt(n)))+A1*X1+A2*X2+A3*X3+A4*X4+A5*X5
      +rnorm(length(X1),0,1) #30
      #a=A0+(A1+(lam*sig/sqrt(sx1)))*X1+A2*X2+A3*X3+A4*X4+A5*X5
      +rnorm(length(X1),0,1) #31
      #a=A0+A1*X1+(A2+(lam*sig/sqrt(sx2)))*X2+A3*X3+A4*X4+A5*X5
      +rnorm(length(X1),0,1) #32
      #a=A0+A1*X1+A2*X2+A3*X3+A4*X4+A5*X5+rnorm(length(X1),0,lam*sig) #33
      b=A0+A1*X1+A2*X2+A3*X3+A4*X4+A5*X5+rnorm(length(X1),0,1) #34
      alea=sample(k,ml) #35
    }
  }
}

```

```

#Y=if(alea[1]==i||alea[2]==i) a else b #36
#Y=if(alea[1]==i||alea[2]==i||alea[3]==i||alea[4]==i) a else b #37
Y=if(alea[1]==i||alea[2]==i||alea[3]==i||alea[4]==i||alea[5]==i)
a else b #38
#Y=if(alea[1]==i||alea[2]==i||alea[3]==i||alea[4]==i||alea[5]==i
||alea[6]==i)a else b #39
#Y=if(alea[1]==i||alea[2]==i||alea[3]==i||alea[4]==i||alea[5]==i
||alea[6]==i||alea[7]==i||alea[8]==i||alea[9]==i||alea[10]==i)
a else b #40

#Y<-if(i<ml) b else a #41

reg=lm(Y ~ X1 + X2 + X3 + X4 + X5) #42
CR=coefficients(reg) #43
datos[i,]=CR #44
}

s=cov(datos) #45
m=colMeans(datos) #46
#s=cov.mve(datos)$cov #47
#m=cov.mve(datos)$center #48
#s=covMcd(datos)$raw.cov #49
#m=covMcd(datos)$raw.center #50
#s=covMcd(datos,alpha=0.5)$cov #51
#m=covMcd(datos,alpha=0.5)$center #52
#s=covMcd(datos,alpha=0.75)$cov #53
#m=covMcd(datos,alpha=0.75)$center #54

sinv=solve(s) #55

T2=matrix(0,k,1) #56

for(j in 1:k) #57
{
  df=datos[j,]-m #58
  df=as.matrix(df) #59
  T=df%*%sinv%*%t(df) #60
  T2[j,]=T #61
}

Max=max(T2) #62
MX[g,]=Max #63
if(Max>LCS) #64
PS[g,]=1 #65
}

mean(PS) #66
}

sñ=SEÑAL(X1,X2,X3,X4,X5,Ao,A1,A2,A3,A4,A5,k,LCS,ml,sig,lam,p) #67
sñ #68

lam<-seq(0.5,5,0.5) #69
SÑ<-matrix(0,length(lam),1) #70

```

```

for(h in 1:length(lam)) #71
{
sñ<-SEÑAL(X1,X2,Ao,A1,A2,k,LCS,ml,sig,lam[h]) #72
SÑ[h,]<-sñ #73
}
SÑ #74

```

Para correr este programa se necesita los siguientes parámetros de entrada

#1-#6: Un intercepto y cinco constantes asociadas a cada variable respectivamente

#7-#11: Cinco variables independientes

#12: n: Tamaño de las variables

#13-#15: Constantes asociadas a los cambios en los parámetros

#16: k: Número de muestras para cada iteración

#17: l: Número de iteraciones

#18: LCS: Límite de control superior

#19: ml: Número de contaminaciones

#20: p: Número de variables

El programa funciona de la siguiente manera

#21: Se carga la libreria MASS

#22: Se carga la libreria rrcov

#23: Se crea una función llamada SEÑAL que depende de 17 parámetros

#24: Se crea una matriz MX de tamaño $l \times 1$ que guardará el valor de la máxima estadística T2 en cada iteración

#25: Se crea una matriz PS de tamaño $l \times 1$ que guardará el valor 1 si la máxima estadística T2 se sale fuera del límite de control en cada iteración

#26: Se crea una for para generar 10000 iteraciones

#27: En cada iteración se crea una matriz datos de tamaño $k \times p+1$ que guardara los coeficientes de regresión para cada muestra o perfil

#28: Se crea un data.frame

#29: Se crea una for para generar k muestras en cada iteración

#30-#33: Se contamina el intercepto, el parámetro asociado a la variable X1, el parámetro asociado a la variable X2 y la varianza en cada iteración. Se debe elegir entre una de estas contaminaciones para la simulación.

#34: Para cada muestra se genera un modelo Y bajo control con los parámetro de entrada

#35: De las k muestras se genera un número aleatorio para contaminar ml muestras

#36-#40: De las k muestras se genera ml perfiles fuera de control y los restantes en control y dependiendo del número deseado de contaminaciones aleatorias, se elige una de las líneas

-
- #41: En el caso de realizar la contaminación a través de un cambio sostenido se debe elegir esta línea y se contaminarán los últimos ml perfiles
- #42-#43: Se estiman los parámetros asociados al modelo
- #44: Para cada muestra se almacenan los coeficientes de regresión del perfil en datos
- #45-#54: Se elige el estimador deseado y se estima la media y la matriz de varianzas asociadas a las k muestras
- #55: Se calcula la inversa de la matriz s
- #56: Se crea una matriz T2 de tamaño k x 1 que guardará el valor de la estadística T2 para cada muestra
- #57: Se crea una for para generar los valor de la estadística T2 para cada muestra
- #58-#60: Se calcula la estadística T2 para cada muestra
- #61: Se almacena la estadística T2 para cada muestra en la matriz T2
- #62: Se calcula el valor máximo de la estadística T2 para las k muestras
- #63: Se almacena el valor máximo en la matriz MX
- #64-#65: Si el valor máximo de la estadística T2 es mayor al límite de control se guarda 1 en la matriz PS
- #66: Se calcula la probabilidad de una señal fuera de control
- #67-#68: Se llama la función sñ
- #69-#74: Se crea la función SÑ para calcular cada una de las probabilidades cambiando el parámetro lamda

Conclusiones

- Cuando los estimadores de localización y dispersión usuales se usan en el análisis de perfiles de regresión lineal múltiple en Fase I, no es fácil detectar múltiples perfiles atípicos debido a un efecto conocido como enmascaramiento. En general, el método Usual presenta un desempeño muy pobre en comparación con los métodos robustos.
- El estudio de simulación mostró que para pocos perfiles atípicos el mejor método para detectar la probabilidad de una señal fuera de control es RMCD75, para una cantidad mayor de perfiles atípicos los métodos MVE y RMCD50 presentan el mejor desempeño.
- En el caso de un cambio sostenido en los parámetros del modelo, las cartas de control robustas en general son muy eficientes para detectar cambios en los coeficientes de regresión asociados a cada variable cuando se presentan pocas contaminaciones, excepto por la carta RMCD75 que presenta un desempeño muy pobre. En particular los métodos MVE y RMCD50 presentan el mejor desempeño.
- Para altas contaminaciones cuando se genera cambios sostenidos en los parámetros del modelo, ningún método presenta un buen desempeño.
- El ejemplo aplicado a un conjunto de datos reales mostró que las cartas de control robustas en general son muy eficientes para detectar perfiles atípicos. La carta usual es ineficiente.

Trabajo futuro

- Cuando el proceso está fuera de control es importante identificar las causas de la anomalía, con el fin de aplicar las medidas correctivas apropiadas. Un trabajo futuro puede implementar herramientas de diagnóstico que permitan determinar cuáles son los parámetros responsables de una señal fuera de control.
- El interés principal en este trabajo fue monitorear simultáneamente el intercepto y los parámetros asociados a cada variable independiente. Sin embargo, no se monitoreó la varianza de los perfiles. Un trabajo futuro podría incorporar cartas de control basadas en estimadores robustos para monitorear la varianza del proceso en Fase I.
- Durante este trabajo, se ha asumido que los valores X son constantes conocidas y toman el mismo conjunto de valores para cada muestra. Si los valores X son aleatorios, entonces bajo ciertas condiciones se puede usar los métodos para Fase I simplemente al modificar las fórmulas para representar la variación de los valores X de muestra a muestra (Mahmoud et al. (2007)). Un trabajo futuro podría abordar esta modificación.
- Los ejercicios de simulación se basaron en un número fijo de observaciones para cada perfil ($n = 10$). Un estudio futuro podría realizar simulaciones para valores diferentes de n con el objetivo de determinar si estas variaciones modifican los resultados obtenidos.
- Los estimadores robustos que se propusieron en este trabajo, conducen a una probabilidad baja para detectar cambios pequeños en el intercepto en Y . En un trabajo futuro se recomienda usar por ejemplo, la carta de control T^2 sugerida por Sullivan & Woodall (1996), ya que tal como lo mostró Vargas (2003), es eficiente para detectar cambios de nivel en el vector de medias.

Glosario

EWMA	Promedios móviles ponderados exponencialmente
ISO	Organización Internacional de Normalización
LCS	Límite de control superior
MCD	Determinante de covarianza mínima
MVE	Elipsoide de volumen mínimo
RMCD50	Determinante de covarianza mínima reponderado con punto de ruptura del 50 %
RMCD75	Determinante de covarianza mínima reponderado con punto de ruptura del 25 %
SPC	Control estadístico de procesos
T_{MCD}^2	Carta T^2 basada en los estimadores del Determinante de Covarianza Mínima
T_{MVE}^2	Carta T^2 basada en los estimadores del Elipsoide de Volumen Mínimo
T_{RMCD50}^2	Carta T^2 basada en los estimadores del Determinante de Covarianza Mínima Reponderado con punto de ruptura del 50 %
T_{RMCD75}^2	Carta T^2 basada en los estimadores del Determinante de Covarianza Mínima Reponderado con punto de ruptura del 25 %
T_{Usual}^2	Carta T^2 basada en el vector de medias muestral y la matriz de covarianzas muestral

Bibliografía

- [1] ALT, F. B. (1984), *Multivariate Quality Control*, Encyclopedia of Statistical Sciences, John Wiley, New York, 110-122.
- [2] ALT, F. B. & SMITH, N. D. (1988), *Multivariate Process Control*, Handbook of Statistics, P. R. Krishnaiah and C. R. Rao, North-Holland, Amsterdam, 7, 333-351.
- [3] CAMPBELL, N. A. (1980), *Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation*, Applied Statistics, 29, 231-237.
- [4] CHENOURI, S. VARIYATH, A. M. & STEINER, S. H. (2009), *A Multivariate Robust Control Chart for Individual Observations*, Journal of Quality Technology, 41:3, 259-271.
- [5] CROARKIN, C. & VARNER, R. (1982), *Measurement Assurance for Dimensional Measurements on Integrated-Circuit Photomasks*, NBS Technical Note 1164. U.S. Department of Commerce, Washington D.C.
- [6] CROUX, C. & HAESBROECK, G. (1999), *Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator*, Journal of Multivariate Analysis, 71, 161-190.
- [7] DAVIES, P. L. (1987), *Asymptotic Behavior of S-Estimates of Multivariate Location Parameters and Dispersion Matrices*, The Annals of Statistics, 15, 1269-1292.
- [8] DEMING, W. E. (1982), *Out of the Crisis*, MIT Press, Cambridge.
- [9] DONOHO, D. L. (1982), *Breakdown Properties of Multivariate Location Estimators*. Ph.D. Qualifying Paper, Harvard University.
- [10] DONOHO, D. L. & HUBER, P. J. (1983), *The Notion of Breakdown Point*. In A Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday, P. J. Bickel, K. A. Doksum, and J. L. Hodges, Jr., eds., 157-184. Belmont, CA: Wadsworth.
- [11] GULLIKSEN, H. & WILKS, S. S. (1950), *Regression test for several samples*, Psychometrika, 15, 91-114.
- [12] GUPTA, S. MONTGOMERY, D. C. & WOODALL, W. H. (2006), *Performance evaluation of two methods for online monitoring of linear calibration profiles*, International Journal of Production Research, 44, 1927-1942.

-
- [13] HARDIN, J. & ROCKE, D. M. (2004), *Outlier Detection in the Multiple Cluster Setting Using the Minimum Covariance Determinant Estimator*, Computational Statistics and Data Analysis, 44, 625-638.
- [14] HARDIN, J. & ROCKE, D. M. (2005), *The Distribution of Robust Distances*, Journal of Computational and Graphical Statistics, 14, 928-946.
- [15] HAWKINS, D. M. & OLIVE, D. J. (1999), *Improved Feasible Solution Algorithm for High Breakdown Estimation*, Computational Statistics and Data Analysis, 30, 1-11.
- [16] HOTELLING, H. (1947), *In Techniques of Statistical Analysis*, New York, McGraw-Hill, 111-184.
- [17] HUBER, P. J. (1981), *Robust Statistics*, John Wiley & Sons, New York, NY.
- [18] JENSEN, W. A. BIRCH, J. B. & WOODALL, W. H. (2007), *High Breakdown Estimation Methods for Phase I Multivariate Control Charts*, Quality and Reliability Engineering International, 23:5, 615-629.
- [19] KANG, L. & ALBIN, S. L. (2000), *On-Line Monitoring When the Process Yields a Linear Profile*, Journal of Quality Technology, 32, 418-426.
- [20] KIM, K. MAHMOUD, M. A. & WOODALL, W. H. (2003), *On the Monitoring of Linear Profiles*, Journal of Quality Technology, 35, 317-328.
- [21] LOPUHAÄ, H. P. (1989), *On the Relation Between S-Estimators and M-Estimator of Multivariate Location and Covariance*, The Annals of Statistics, 17, 1662-1683.
- [22] LOPUHAÄ, H. P. & ROUSSEEUW, P. J. (1991), *Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices*, The Annals of Statistics, 19, 229-248.
- [23] MAHMOUD, M. A. & WOODALL, W. H. (2004), *Phase I Monitoring of Linear Profiles with Calibration Application*, Technometrics, 46, 380-391.
- [24] MAHMOUD, M. A. PARKER, P. A. WOODALL, W. H. & HAWKINS, D. M. (2007), *A Change Point Method for Linear Profile Data*, Quality and Reliability Engineering International, 23, 247-268.
- [25] MAHMOUD, M. A. (2008), *Phase I Analysis of Multiple Linear Regression Profiles*, Communications in Statistics - Theory and Methods, 37:10, 2106-2130.
- [26] MARONNA, R. A. (1976), *Robust M-Estimators of Multivariate Location and Scatter*, Annals and Statistics, 4, 51-67.
- [27] MESTEK, O. PAVLIK, J. & SUCHANEK, M. (1994), *Multivariate Control Charts: Control Charts for Calibration Curves*, Fresenius Journal of Analytical Chemistry, 350, 344-351.
- [28] MONTGOMERY, D. C. (2009), *Introduction to Statistical Quality Control*, 6th edn, John Wiley and Sons, New York.
- [29] MYERS, R. H. (1990), *Classical and Modern Regression with Applications*, 2nd edn, PWS-Kent Publishing Company, Boston.

-
- [30] PEÑA, D. (2002), *Análisis de Datos Multivariantes Manuscrito*, Universidad Carlos III de Madrid.
- [31] PISON, G. & VAN ALEST, S. & WILLEMS, G. (2002), *Small Sample Corrections for LTS and MCD*, *Metrika*, 55, 111-123.
- [32] ROUSSEEUW, P. J. (1984), *Least Median of Squares Regression*, *Journal of American Statistical Association*, 79, 871-880.
- [33] ROUSSEEUW, P. J. (1985), *Multivariate Estimation with High Breakdown Point*. In *Mathematical Statistics and Applications*, Section B. W. Grossmann, G. Pflug, I. Vincze, and W. Werz, eds., 283-297. Dordrecht: Reidel.
- [34] ROUSSEEUW, P. J. & LEROY, A. M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York, NY.
- [35] ROUSSEEUW, P. J. & VAN DRIESSEN, K. (1999), *A Fast Algorithm for the Minimum Covariance Determinant Estimator*, *Technometrics*, 41, 212-223.
- [36] ROUSSEEUW, P. J. & YOHAI, A. M. (1984), *Robust Regression by Means of S-Estimators*, In *Robust and Nonlinear Time Series Analysis*. *Lecture Notes in Statistics* 26, 256-272.
- [37] ROUSSEEUW, P. J. & VAN ZOMEREN, B. C. (1990), *Unmasking Multivariate Outliers and Leverage Points*, *Journal of American Statistical Association*, 85, 633-651.
- [38] SERFLING, R. (1980), *Approximation Theorems of Mathematical Statistics*. New York, NY: John Wiley and Sons.
- [39] STAHEL, W. A. (1981), *Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators*. Ph.D. Dissertation, ETH, Zurich (in German).
- [40] SULLIVAN, J. H. & WOODALL, W. H. (1996), *A Comparison of Multivariate Control Charts for Individual Observations*, *Journal of Quality Technology*, 28, 398-408.
- [41] STOVER, F. S. & BRILL, R. V. (1998), *Statistical Quality Control Applied to Ion Chromatography Calibrations*, *Journal of Chromatography*, A804, 37-43.
- [42] TRACY, N. D. YOUNG, J. C. & MASON, R. L. (1992), *Multivariate Control Charts for Individual Observations*, *Journal of Quality Technology*, 24, 88-95.
- [43] VARGAS, N. J. A. (2003), *Robust Estimation in Multivariate Control Charts for Individual Observations*, *Journal of Quality Technology*, 35, 367-376.
- [44] VARGAS, N. J. A. (2006), *Control Estadístico de Calidad*, Universidad Nacional de Colombia.
- [45] VARIYATH, A. M. & VATTATHOOR, J. (2013), *Robust Control Charts for Monitoring Process Variability in Phase I Multivariate Individual Observations*, *Quality and Reliability Engineering International*.
- [46] WILLEMS, G. PISON, G. ROUSSEEUW, P. J. & VAN ALEST, S. (2002), *A Robust Hotelling Test*, *Metrika*, 55, 125-138.

-
- [47] WOODALL, W. H. (2000), *Controversies and Contradictions in Statistical Process Control* (with discussion), *Journal of Quality Technology*, 32, 341-378.
- [48] WOODALL, W. H. SPITZNER, D. J. MONTGOMERY, D. C. & GUPTA, S. (2004), *Using Control Charts to Monitor Process and Product Profiles*, *Journal of Quality Technology*, 36, 309-320.
- [49] WOODALL, W. H. (2007), *Current Research in Profile Monitoring*, *Revista Producao*, 17(3), 420-425.
- [50] WOODRUFF, D. L. & ROCKE, D. M. (1994), *Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators*, *Journal of American Statistical Association*, 89, 888-896.
- [51] ZOU, C. TSUNG, F. & WANG, Z. (2006A), *Monitoring General Linear Profiles Using Multivariate EWMA Schemes*, *Technometrics*, 50:4, 512-526.
- [52] ZOU, C. ZHANG, Y. & WANG, Z. (2006B), *A Control Chart Based on a Change-Point Model for Monitoring Profiles*, *IIE Transactions*, 38, 1093-1103.