

# Taller 1 - Análisis de datos funcionales

Valeria Bejarano, Camilo Avellaneda

## Contents

1 Desarrollo	3
--------------	---

1. El proceso de producción de azúcar proveniente de la remolacha es muy importante en ciertas regiones del mundo como en Escandinavia. En este proceso se emplea la espectrometría para detectar impurezas y controlar la calidad del azúcar. En particular las curvas espectrométricas corresponden a 7 longitudes de onda de excitación ( $Y_1 = 230, Y_2 = 240, Y_3 = 255, Y_4 = 290, Y_5 = 305, Y_6 = 325$  y  $Y_7 = 340\text{nm}$ ), donde cada una de estas longitudes representa un proceso funcional continuo ( $Y_i$ ), así que el proceso multivariado funcional  $\mathbf{Y}$  está conformado por  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_7)$ .

Usando este dataset seleccione un proceso cualesquiera y para este proceso encuentre

- La función media
- La función media recortada al 10%
- La función varianza
- La función covarianza
- La función mediana
- Funciones cuantiles 90 y 95
- Región central 0.75
- Outliers
- El factor de expansión de 0.5 central región,  $F$ , tal que la tasa de falsos outliers sea de 0.007.

2. Usando este mismo dataset, encuentre:

- La función mediana multivariada
- Los outliers multivariados.
- Encuentre la correlación entre los diferentes procesos usando los coeficientes de correlación de Kendall y de Spearman para datos funcionales.

3. No existe una única manera de determinar la profundidad de objetos funcionales, de hecho, en la literatura existe una gran cantidad de propuestas. De acuerdo a la siguiente distribución, estudie la propuesta Extremal depth e impleméntela usando el conjunto de datos asociado a la producción de azúcar.
4. De acuerdo a la siguiente distribución, estudie la propuesta Dai, W., & Genton, M. G. (2019). Directional outlyingness for multivariate functional data. Computational Statistics & Data Analysis, 131, 50-65. e impleméntela usando el conjunto de datos asociado a la producción de azúcar.

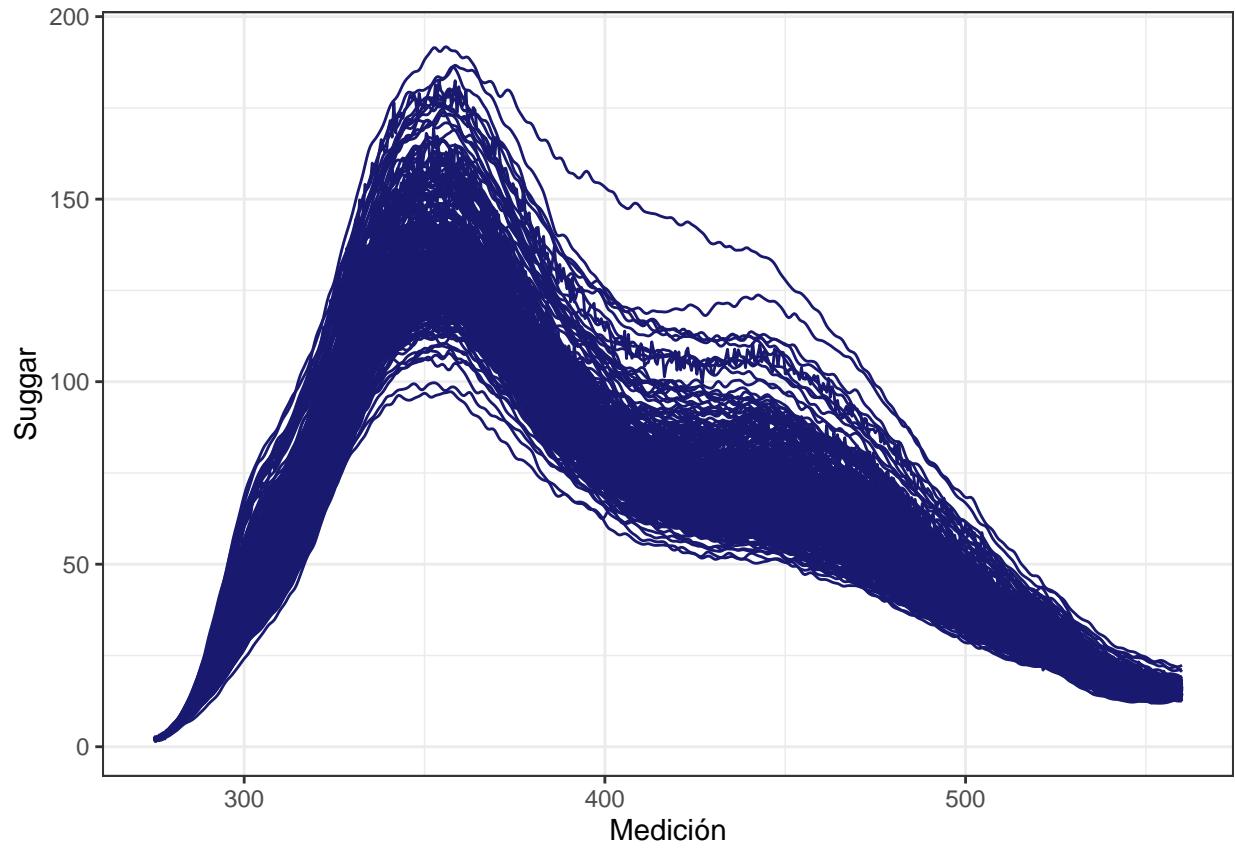


Figure 1: Gráfico de curvas suavizadas de las 268 funciones tenidas en cuenta dentro de la categoría de la longitud considerada.

# 1 Desarrollo

En este trabajo se busca explorar las diferentes herramientas dadas para el análisis de datos funcionales sobre un proceso de producción de azúcar. Para siete longitudes de onda de excitación diferentes se tienen 268 curvas espectrométricas. El trabajo aborda conceptos como el cálculo de una media, desviación, varianza, cuartiles y percentiles en el contexto de datos funcionales. El software utilizado para todos los procedimientos es R [ver R Core Team, 2020]. La exploración de este conjunto de datos funcionales se realiza inicialmente desde una perspectiva univariada, en la cual consideramos la realización asociada a  $Y_5$ , mientras en una segunda etapa se considera el proceso completo como un proceso funcional multivariado. Los procedimientos que se muestran en este documento se basan en funciones que en su mayoría ya fueron programadas, las cuales se encuentran en las librerías de R como *fda*, *fda.usc* y *roahd* [ver Ramsay et al., 2020, Febrero-Bande and Oviedo de la Fuente [2012] y Ieva et al. [2019], respectivamente]. En los procedimientos que se requiere el uso de profundidades se hace uso de la banda de profundidad modificada o “MBD” por sus siglas en inglés, a menos de que se especifique algo diferente. La teoría correspondiente al cálculo de la MBD para cada función se puede consultar en López-Pintado and Romo [2009].

A partir del proceso funcional considerado en  $Y_5$ , se explora el conjunto de valores dado por la realización del proceso. La figura 1 muestra los valores discretizados representados mediante una curva suavizada de los diferentes puntos. La muestra de datos funcionales corresponde a un total de 268 curvas. Allí se observa un comportamiento creciente en la primer etapa de la curva y un proceso decreciente en una segunda etapa. No se considera el proceso como cíclico, por lo cual las funciones base son *B*-splines.

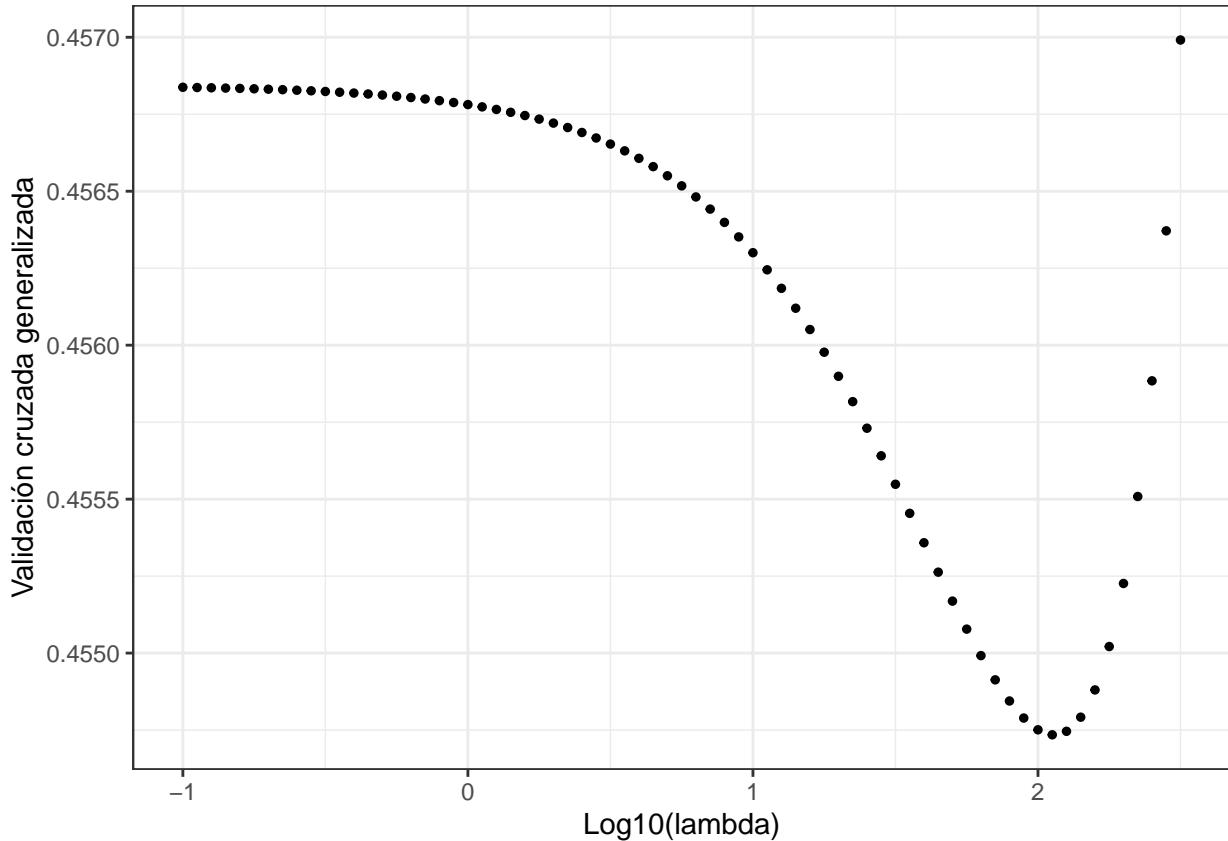


Figure 2: Gráfico de validación cruzada generalizada para determinar el parámetro de penalización óptimo.

Como se muestra en Ramsay and Silverman [2007], la figura 1 ilustra la función de penalización con respecto

al logaritmo del parámetro de curvatura. Entre mayor sea dicho parámetro, las curvas van a tener menos curvaturas y viceversa. En este caso, cuando el logaritmo del parámetro es igual a 2.05, se obtiene un mínimo, por lo cual en adelante se utilizará este valor para todos los procedimientos requeridos.

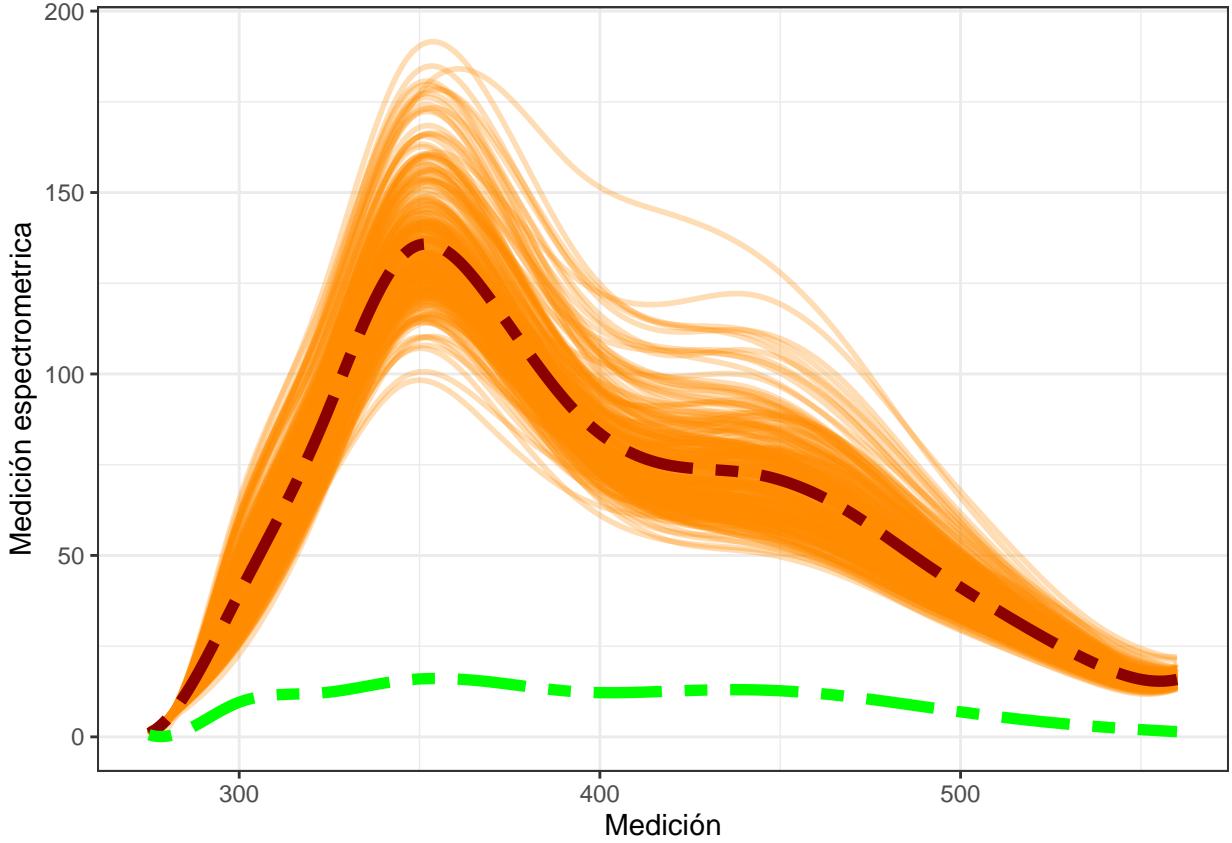


Figure 3: Funciones estimadas junto con su curva promedio y su desviación estándar funcional.

La figura 3 muestra las curvas ajustadas, junto con la curva que representa la función media en color rojo oscuro y la que representa la función asociada a la desviación estándar en color verde.

La figura 4 es una representación análoga a la que se encuentra en la figura 3, pero esta vez considerando la media recortada al 10%. Para ello, se omite el 10% de las curvas con menor profundidad. Este proceso se realizó mediante el argumento *trim*, asociado a la función *mean.fd()*.

Por otro lado, la figura 5 muestra la superficie asociada a la función de varianza y covarianza representada sobre el plano. La diagonal correspondiente a dicha superficie representa la función de varianza, mientras que los valores fuera de la diagonal se asocian con la función de covarianza en dos instantes del tiempo diferentes. La figura 6 muestra la función de varianza y covarianza, pero esta vez representada mediante curvas de nivel. En dichas figuras se observan dos picos. El primero de ellos y más alto se ubica en valores de medición entre 300 y 400, mientras que el segundo se ubica en valores de medición cercanos a 450. En el resto de la figura se observa un comportamiento decreciente en ubicaciones (en el plano cartesiano) diferentes a las mencionadas anteriormente.

Para este documento se consideró la función de la banda de profundidad modificada, las cuales se denotan por “MBD”, por sus siglas en inglés. La figura 7 muestra las 268 funciones y la curva mediana, la cual se obtuvo a partir de las profundidades calculadas. Para ello, se calculó la MBD para las 268 curvas y se seleccionó la mayor.

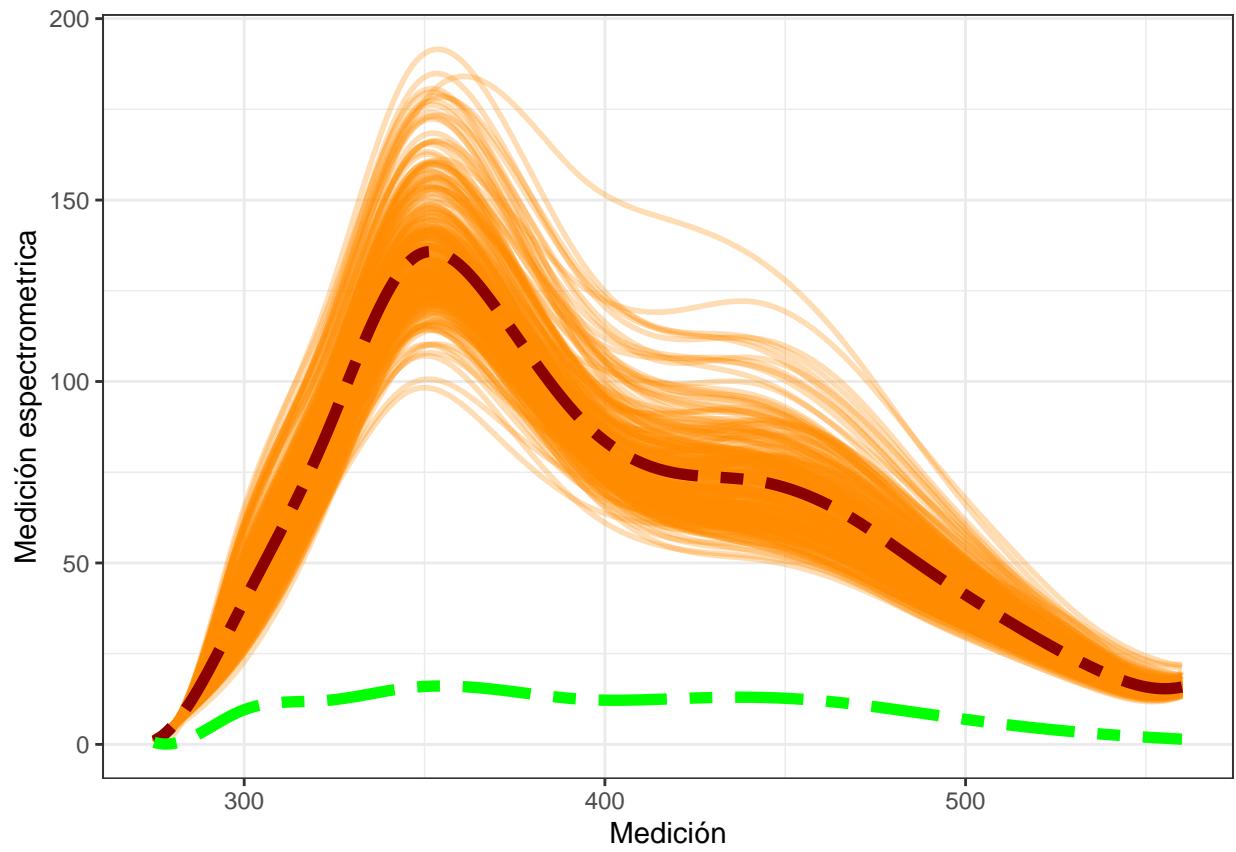


Figure 4: Funciones estimadas junto con su curva promedio recortada al 10% y su desviación estándar funcional

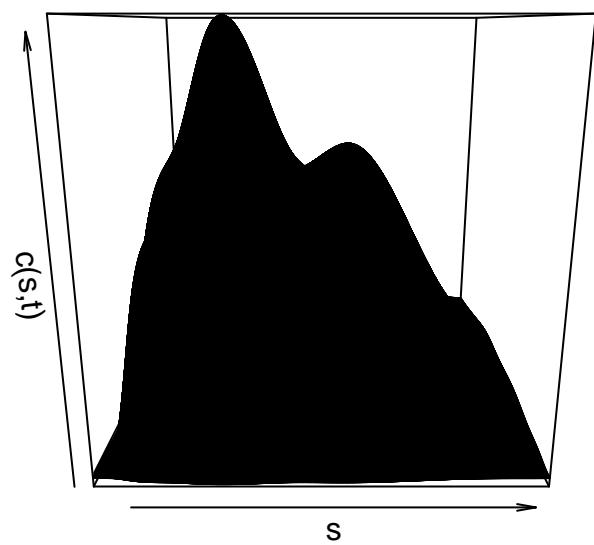


Figure 5: Superficie asociada a la función de varianza y covarianza de la realización obtenida.

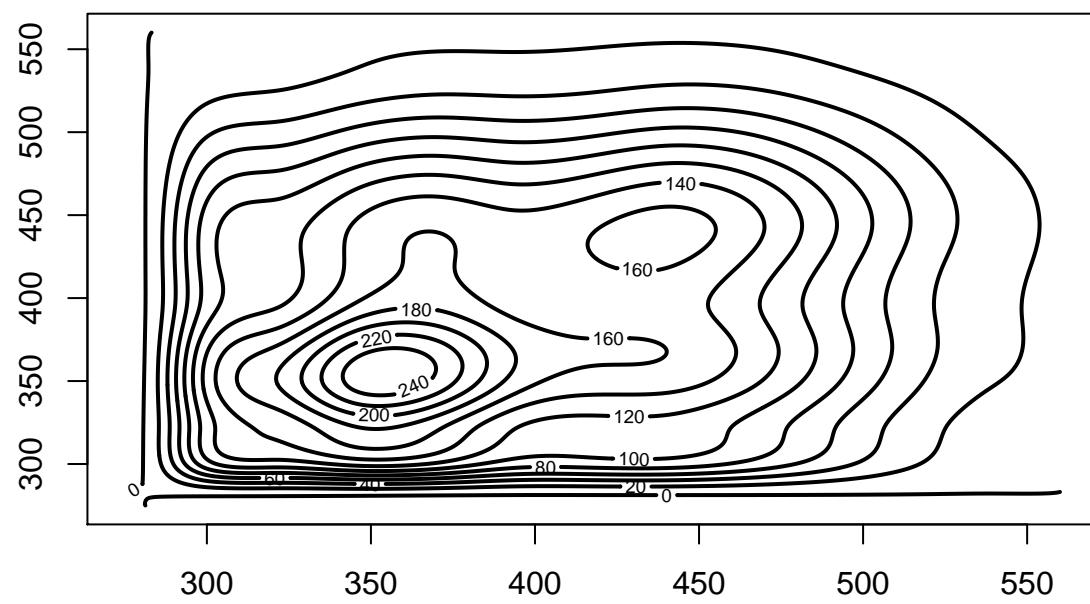


Figure 6: Curvas de nivel calculadas para la función de varianzas y covarianzas.

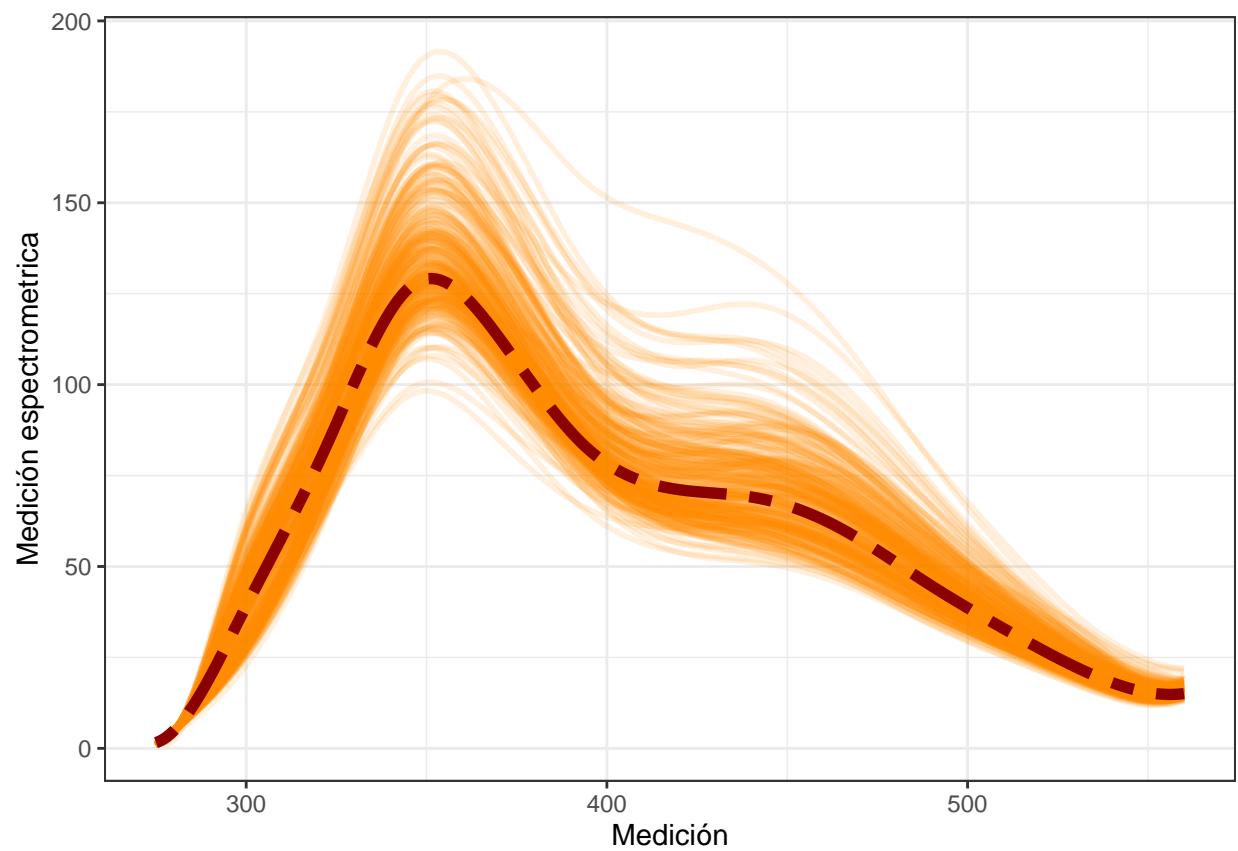


Figure 7: Funciones estimadas junto con la función mediana partir de las profundidades MBD.

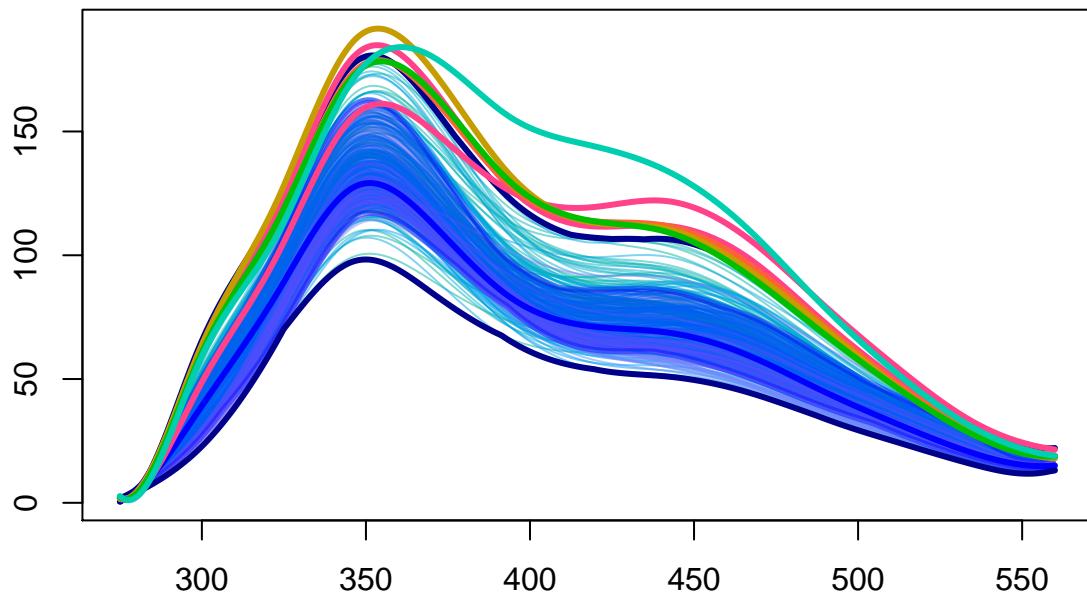


Figure 8: Boxplot funcional univariado para las funciones seleccionadas en Y\_5

La figura @ref(fig:boxplot\_2) muestra el boxplot funcional con una modificación referente a la tasa de outliers permitidos. En este caso se consideró una tasa de outliers de 0.007. Esta modificación va a ampliar las bandas que determinan las curvas que son atípicas y las que no.

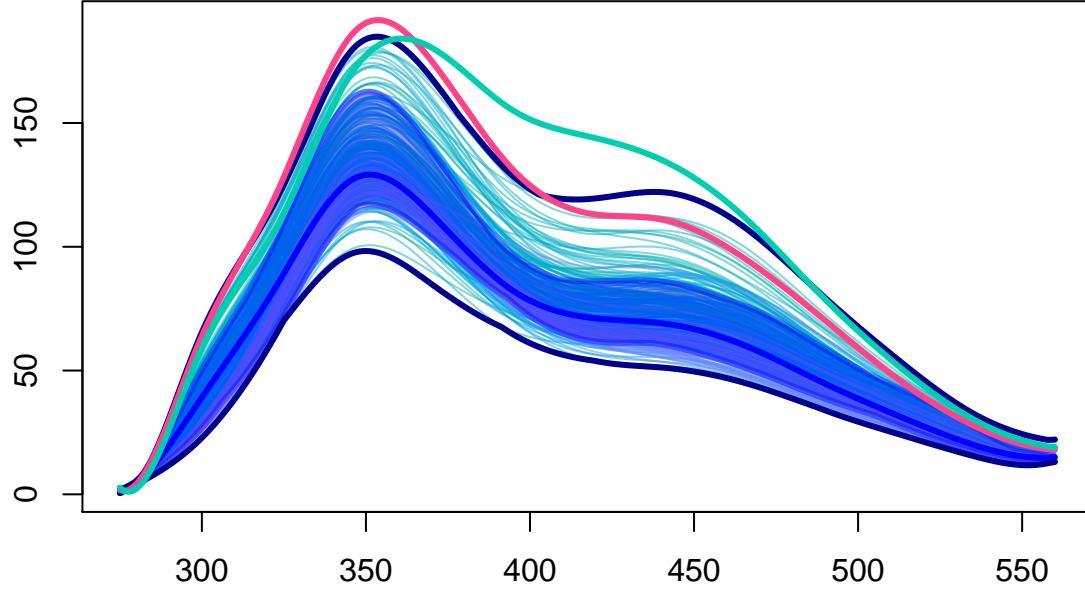


Figure 9: (#fig:boxplot\_2)Boxplot funcional univariado para las funciones seleccionadas en Y\_5 con una tasa de outliers de 0.007

Para determinar si hay puntos outlier se elaboró un boxplot functional, el cual se ilustra en la figura 8, con sus parámetros dados por default para un primer acercamiento. En este gráfico, las curvas azules denotan los límites y los cuartiles funcionales, mientras que las líneas que se presentan en otras tonalidades corresponden a las funciones outliers. Mediante este proceso se encontró un total de 6 curvas atípicas. Los números asociados a éstas curvas son 10, 14, 16, 17, 38, 71.

Para obtener los percentiles funcionales lo que se realiza es encontrar su región central asociada, de forma que al tomar su límite inferior o superior, éste coincide con el percentil deseado. Esto quiere decir que para el percentil 90 lo que se realiza es encontrar las curvas que separan la región central del 80%, para luego tomar el límite superior de esta región central y obtener el percentil deseado. Estos cálculos se realizaron mediante la profundidad MBD. Las figuras 10 y 11 muestran las curvas suavizadas construidas a partir de la realización, además de los percentiles 90 y 95 en líneas punteadas negras, respectivamente. Por otro lado la figura 12 muestra las mismas curvas suavizadas, pero esta vez con dos bandas en color negro que delimitan la región central del 75%.

La figura 13 es una representación de la mediana funcional multivariada. Cada uno de los recuadros en dicha figura corresponde a una de las categorías de las diferentes longitudes, de acuerdo a como se muestra en los respectivos encabezados.

En la figura 14 muestra los boxplot generados a partir de la realización multivariada funcional. Análoga-

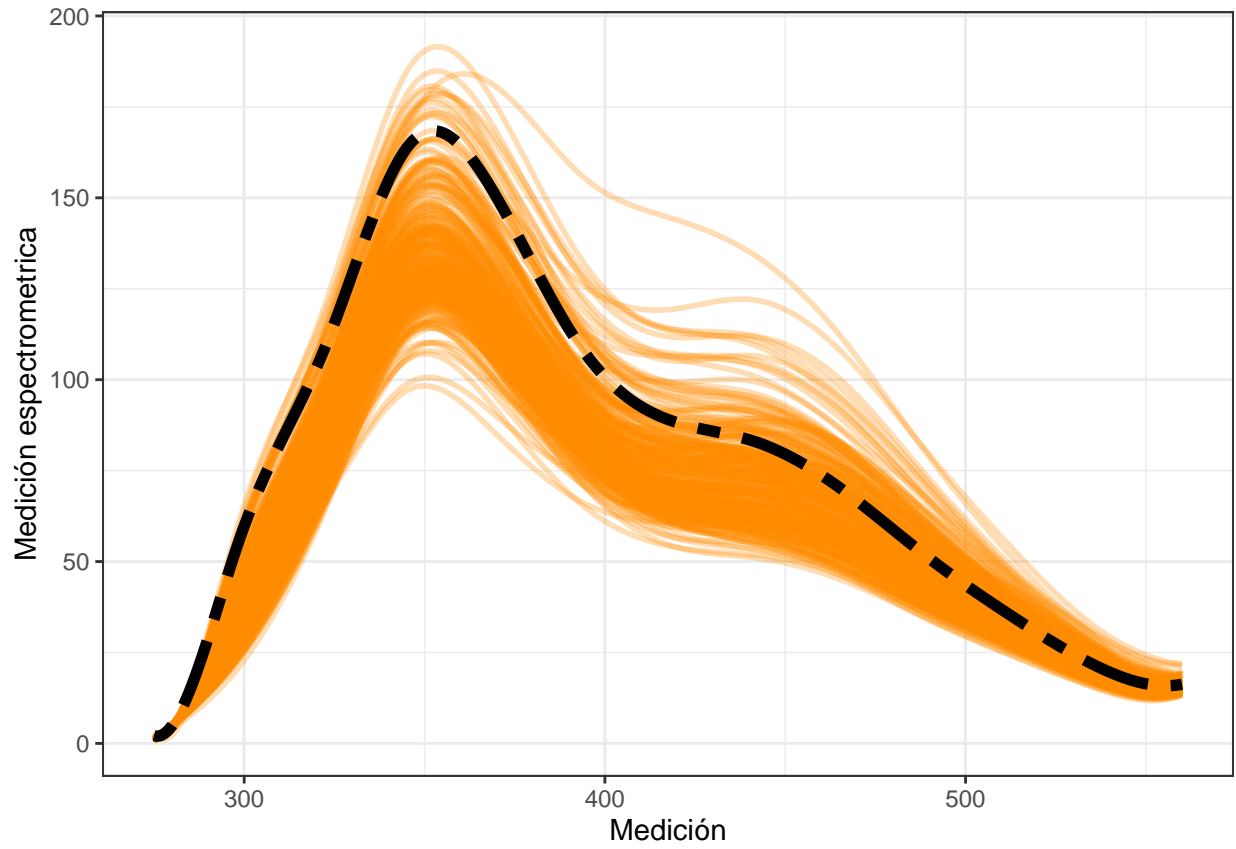


Figure 10: Curvas suavizadas de las mediciones correspondientes a las 268 curvas, junto con su percentil 90, en color negro.

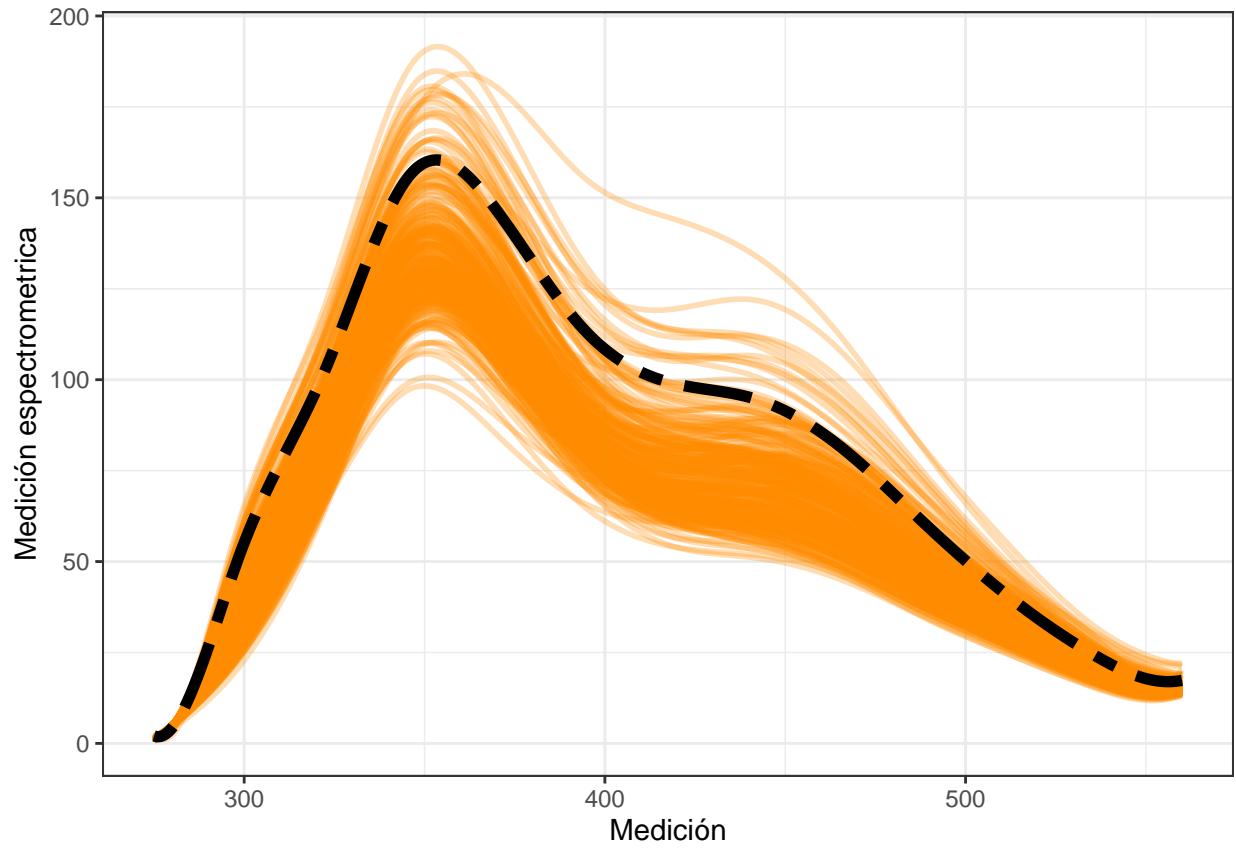


Figure 11: Curvas suavizadas de las mediciones correspondientes a las 268 curvas, junto con su percentil 95, en color negro.

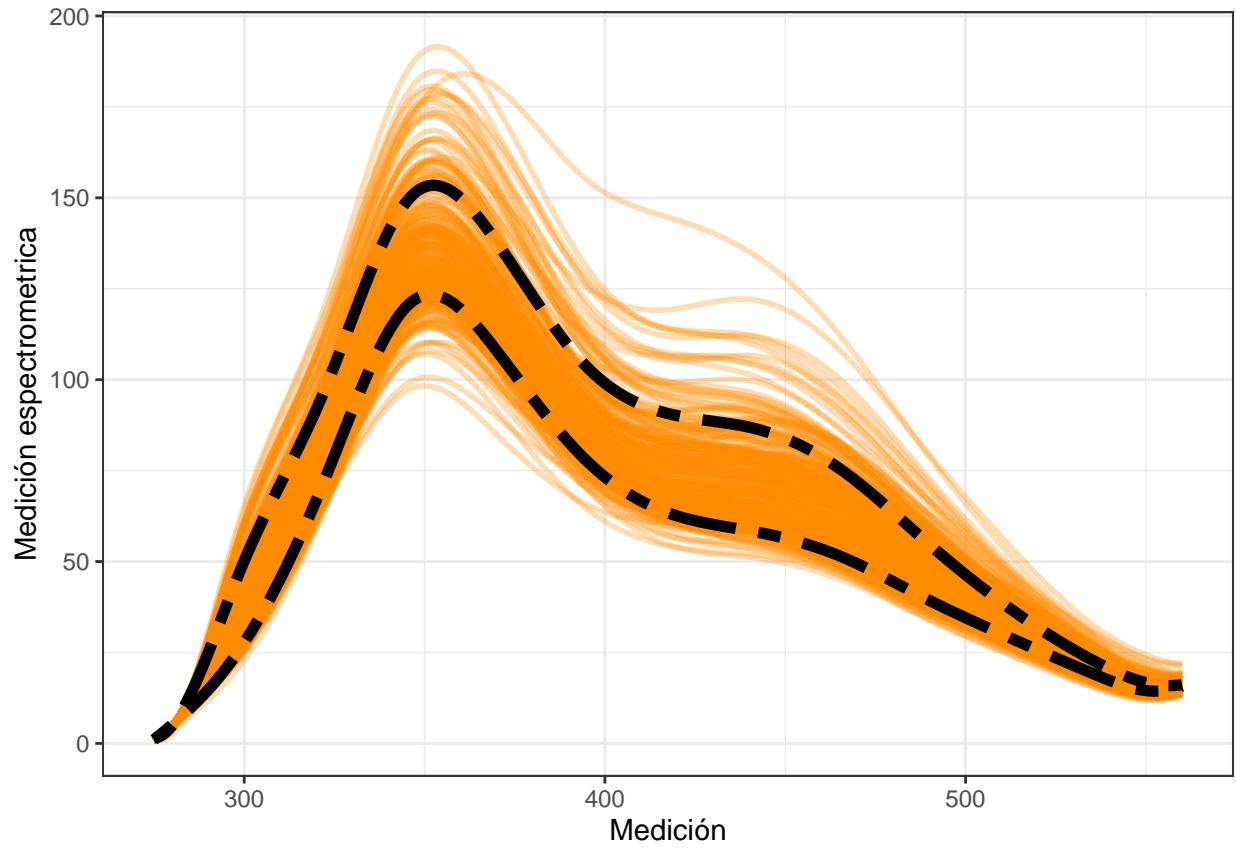


Figure 12: Curvas suavizadas de las mediciones correspondientes a las 268 curvas, junto con la región central de un 75% denotada por las dos bandas en color negro.

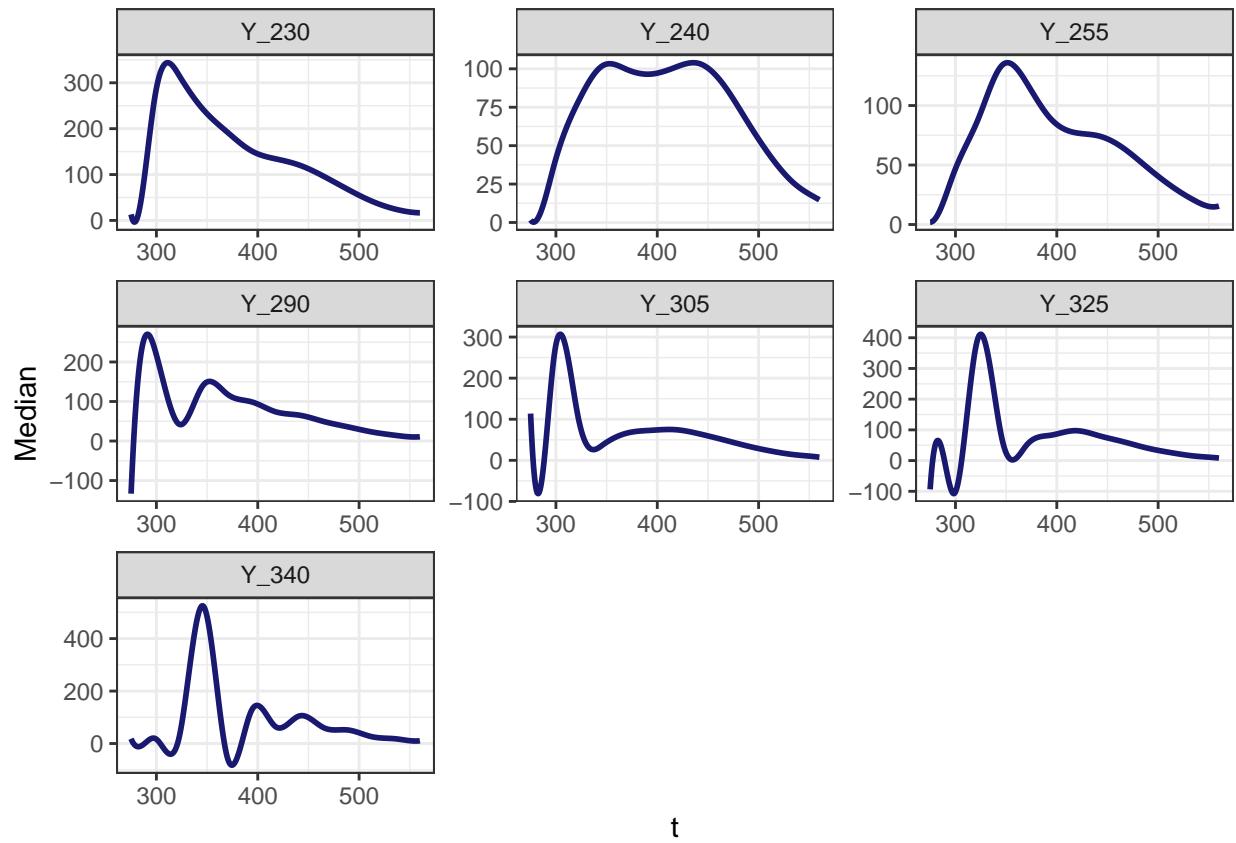


Figure 13: Mediana funcional multivariada correspondiente a la realización considerada.

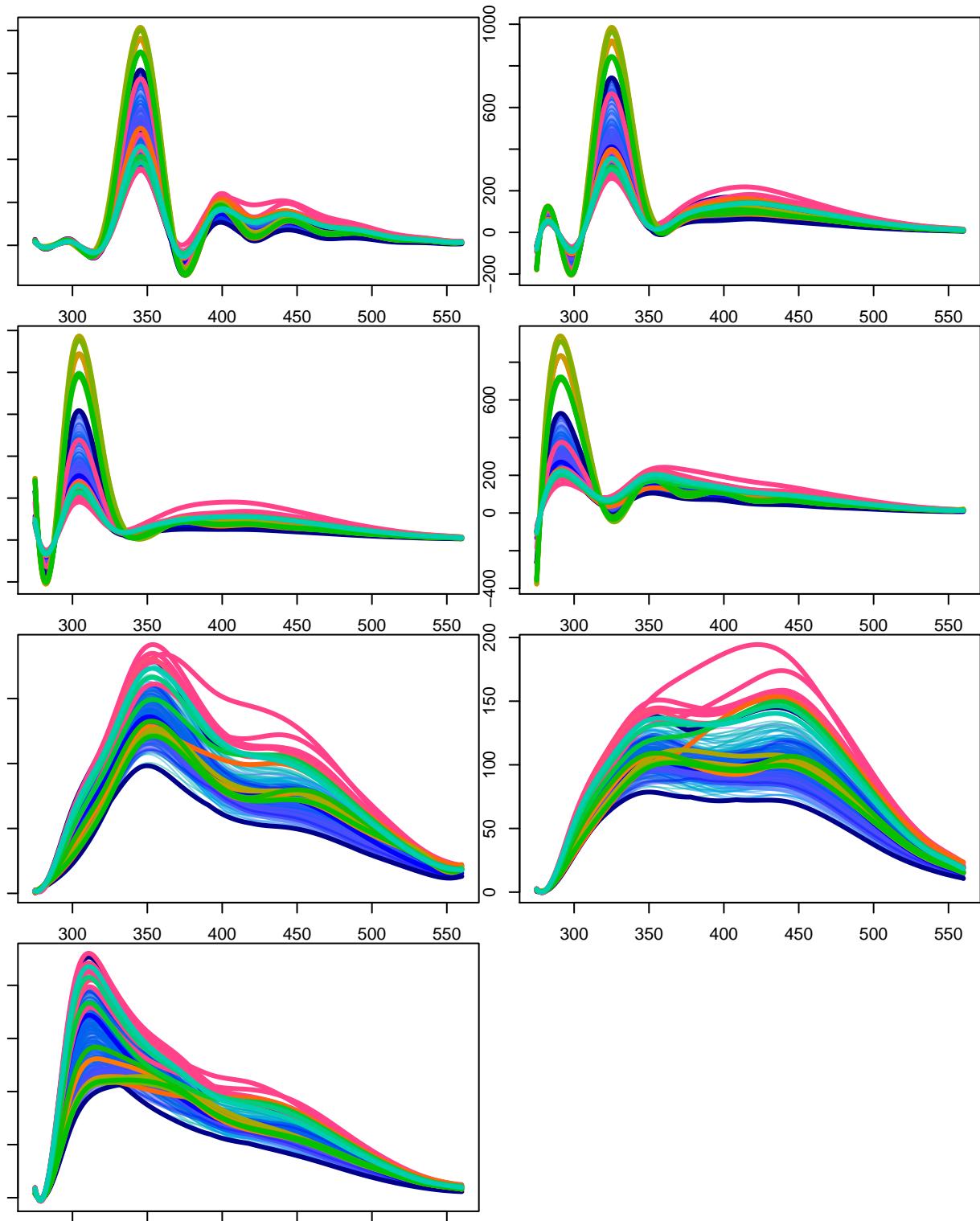


Figure 14: Boxplot funcional multivariado correspondiente a la realización funcional multivariada Y

mente al caso univariado las líneas azules denotan los límites y cuartiles funcionales correspondientes a cada realización funcional por separado, mientras que las líneas en otros colores se asocian con las curvas outliers detectadas por esta metodología. A partir de esta alternativa, se encuentran un total de 18 curvas atípicas y sus códigos son 10, 14, 16, 17, 38, 71, 129, 131, 158, 197, 198, 199, 200, 157, 201, 13, 15, 19.

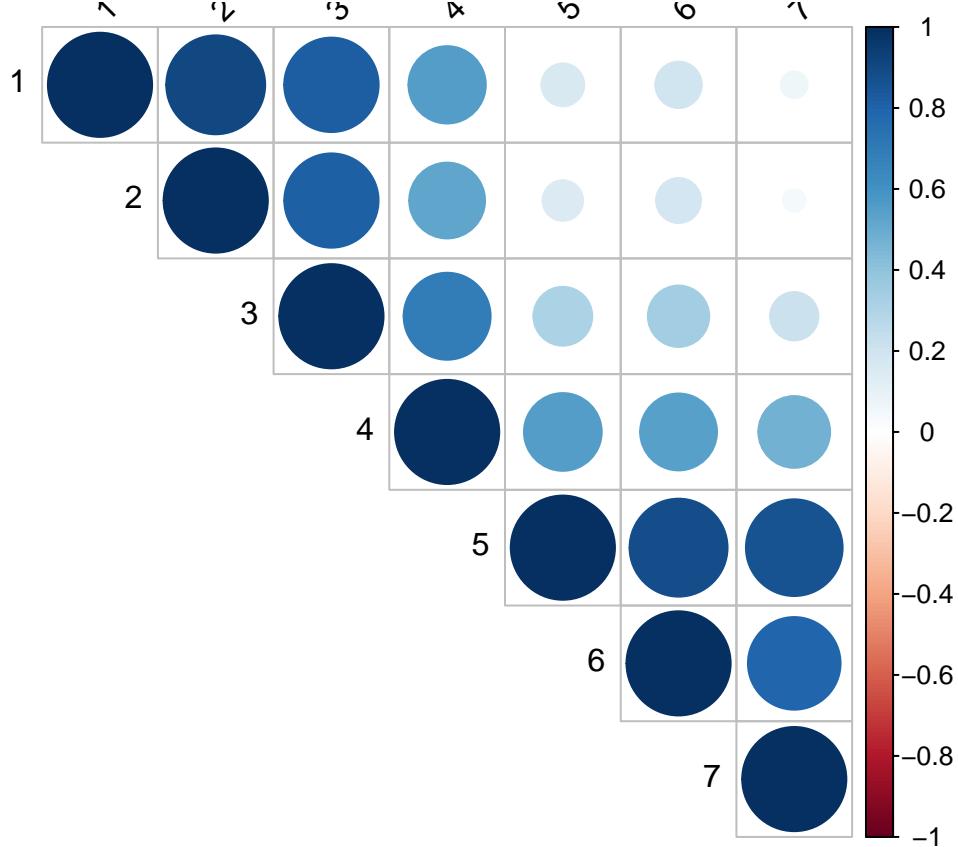


Figure 15: Gráfico de correlaciones de Kendall a partir de la realización multivariada funcional.

La obtención de la correlación de Kendall entre dos procesos funcionales se realiza mediante la concordancia entre curvas, mientras que el coeficiente de Spearman se realiza calculando el coeficiente de correlación de Pearson entre índices generados para cada una de las curvas. La figura 15 es una representación gráfica de una matriz de correlaciones. Allí se observa que la correlación es considerable con el proceso inmediatamente anterior en términos de las longitudes consideradas. La obtención de las correlaciones de Kendall se realizaron mediante el criterio del máximo, lo cual determina si una curva es mayor que otra. Por otro lado, la figura 16 es una representación análoga, pero esta vez a partir de la correlación de Spearman. En este caso, se observa que los procesos funcionales están correlacionados positivamente en su gran mayoría. La teoría correspondiente a las correlaciones de Spearman y de Kendall se puede consultar en Valencia et al. [2014] y Valencia et al. [2019], respectivamente. El cálculo realizado para la correlación de Spearman La obtención de estas correlaciones se realizó mediante las funciones `cor_kendall` y `cor_spearman` de la librería `roahd`.

En la literatura se pueden encontrar diversas propuestas para el cálculo de profundidades. Una de las alternativas es propuesta por Narisetty and Nair [2016] , denotada por profundidad extrema. La motivación para su propuesta se da a partir del hecho de que algunas de las alternativas pueden no ser robustas ante valores atípicos en regiones pequeñas del dominio. Esta propuesta penaliza funciones en intervalos pequeños, incluso si tienen un comportamiento promedio en el resto del dominio. En este artículo, de igual manera se menciona que si se desea caracterizar el comportamiento general de las funciones, otras alternativas para el cálculo de las profundidades de las curvas serían preferibles.

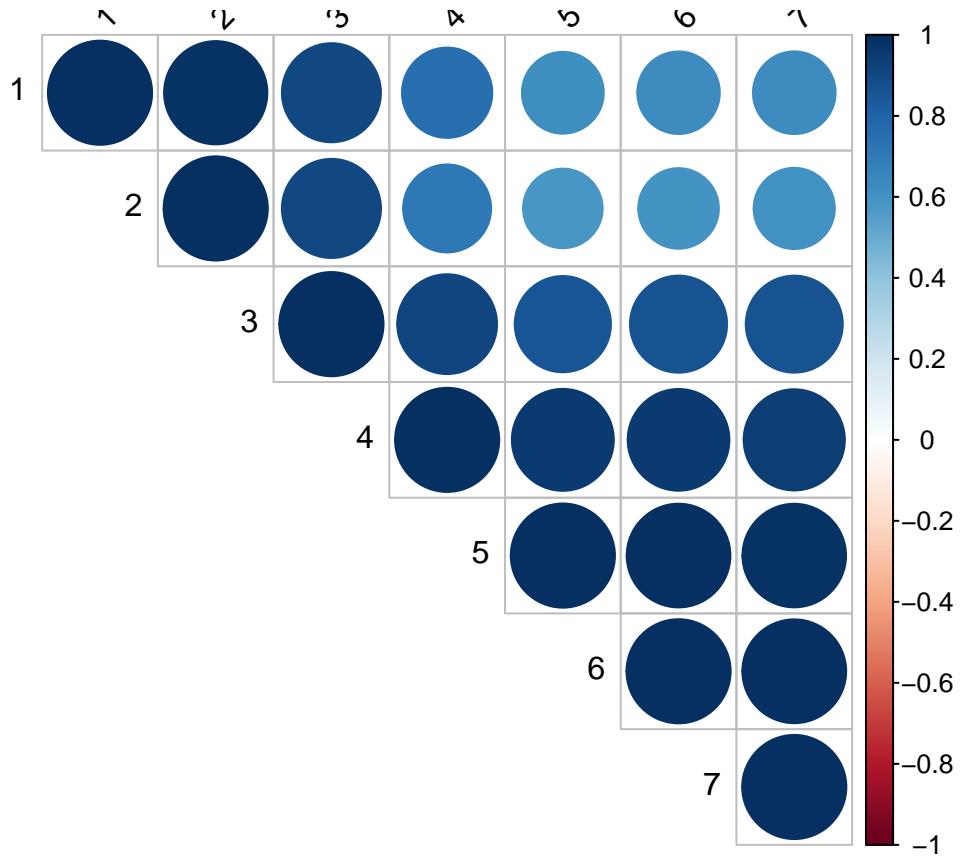


Figure 16: Gráfico de correlaciones de Spearman a partir de la realización multivariada funcional.

El cálculo de la profundidad extrema se realiza como sigue:

Sea  $g(t)$  una función definida sobre el intervalo  $[0, 1]$ . Se define la profundidad punto a punto de  $g(t)$  con respecto a un conjunto de funciones  $S := \{f_1(t), f_2(t), \dots, f_n(t)\}$ , como se muestra en la ecuación (1)

$$D_g(t, S) = 1 - \frac{|\sum_{i=1}^n I_{f_i(t) < g(t)} - I_{f_i(t) > g(t)}|}{n}, \quad (1)$$

donde  $I_A(\cdot)$  representa la función indicadora asociada al evento A. Sea  $\Phi_g(\cdot)$  la función de distribución acumulada de los valores  $D_g(t, S)$  referente a la curva  $g$  en los diferentes tiempos de medición  $t \in [0, 1]$ . En otras palabras, la función  $\Phi_g(\cdot)$  se define como se muestra en la ecuación (2).

$$\Phi_g(r) = \int_0^1 I_{D_g(t, S) < r} dt. \quad (2)$$

La comparación entre dos curvas  $g$  y  $h$  se realiza mediante  $\Phi_g(r)$  y  $\Phi_h(r)$ , de tal forma que se puede determinar si  $g \succ h$ . De esta manera, la profundidad extrema de una función  $g$  con respecto a un conjunto de curvas  $S$  se define como se muestra en la ecuación (3).

$$ED(g, S) = \frac{\sum I_{i: g \succeq f_i}}{n}. \quad (3)$$

La función *extremal\_depth()* del paquete *fdaoutlier* [ver Ojo et al., 2021] calcula las profundidades extremas descritas anteriormente. La figura 17 muestra las diferentes curvas estimadas, donde cada color representa la profundidad extrema calculada para cada una de las curvas.

El artículo de Dai and Genton [2019] nos brinda una herramienta para el cálculo de outliers no solo basados en profundidad si no en una generalización que nos permite su visualización direccional, es decir encontrando outliers tanto de magnitud como de forma, lo que permite un mejor descripción de la centralidad de las curvas y de la variabilidad entre ellas.

La metodología se puede implementar tanto univariada ( $p = 1$ ) como multivariamente ( $p \geq 2$ ), este último caso teniendo presente la correlación de los procesos, que el artículo presenta su eficacia a través de las diferentes simulaciones.

La manera en que se captura tanto la magnitud como la dirección de “outlyingness” es calculando

$$O(X(t), F_{X(t)}) = \{1/d(X(t), F_{X(t)}) - 1\} \cdot \mathbf{v}(t),$$

donde  $X(t)$  es el proceso funcional multivariado,  $F_{X(t)}$  su respectiva función de distribución, que en el caso muestral corresponderá a su correspondiente nube de observaciones,  $d(\cdot)$  una función de profundidad y  $\mathbf{v}(t)$  un vector unitario que apunta de la media del proceso ( $Z(t)$ ) al proceso ( $X(t)$ ).

A partir de esta medida de atipicidad se tiene el cálculo de las siguientes medidas para functional directional outlyingness (FO), mean directional outlyingness (MO) y variation of directional outlyingness (VO), en su versión discreta

$$\begin{aligned} FO(X, F_X) &= 1/n \sum_k \|O(X(t_k), F_{X(t_k)})\|^2 w(t_k) \\ MO(X, F_X) &= 1/n \sum_k O(X(t_k), F_{X(t_k)}) w(t_k) \\ VO(X, F_X) &= 1/n \sum_k \|O(X(t_k), F_{X(t_k)}) - MO(X, F_X)\|^2 w(t_k) \end{aligned}$$

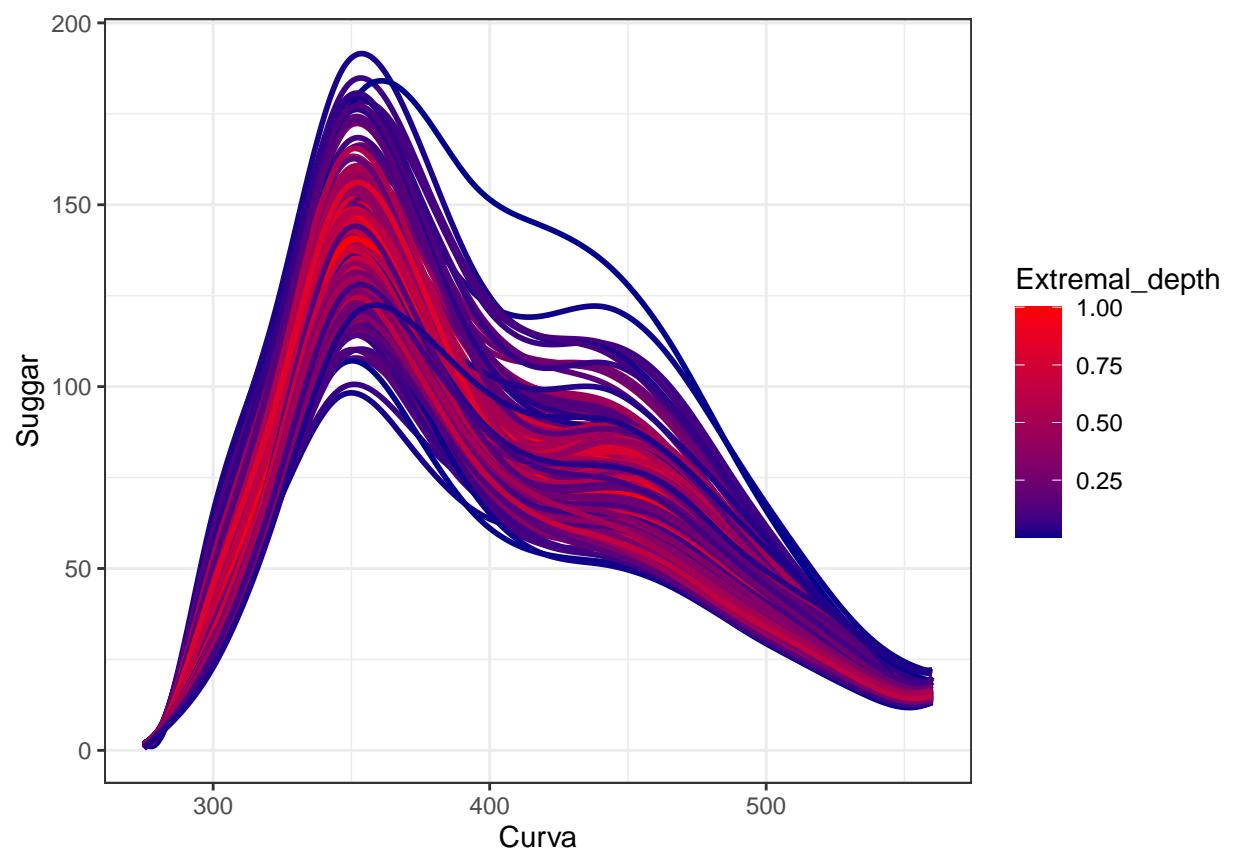


Figure 17: Gráfico de funciones suavizadas, donde el color de la curva representa la profundidad extrema.

Lo interesante es que  $FO(X, F_X) = ||MO(X, F_X)||^2 + VO(X, F_X)$ , es decir la atipicidad total se puede separar en outlier de magnitud ( $MO$ ) y de forma ( $VO$ ) con lo que basta ver quienes tienen estos valores más altos en comparación a la nube, así mediante la función `dir_out` del paquete `fdaoutlier`, el cual usa la profundidad de proyección con base en el outlyingness de Stahel–Donoho ( $(X(t) - \text{Mediana}/MAD)$ ) se obtiene para cada uno de los procesos y de las observaciones un valor  $MO$  y un valor  $VO$  del proceso multivariado para cada observación, en la figura \_\_\_\_\_ se puede ver solo el caso del proceso  $X_3$  en el cual se destacan las curvas 157, 158, 197, 198, 199, 200 y 201 por sus altos valores en  $VO$  lo que nos indica que pueden ser outliers más de forma que de magnitud.

Para el análisis multivariado debido a la proyección en  $\mathbb{R}^p$  en este caso  $p = 7$  no es posible representar en un plano más de 2 procesos junto con el  $VO$ , por eso a manera de exemplificación se obtienen las figuras 19 y 20, para los procesos  $X_1, X_2$  y  $X_4, X_5$  respectivamente. Cabe destacar aquellos puntos que indican un outlier de magnitud más no de escala, el caso de la curva 71 para todos los procesos en estudio, incluso evaluandolo en los procesos 6 y 7 también destaca con un valor grande de  $MO$  más no de  $VO$ , así bajo esta metodología la curva 71 (multivariadamente) es considerada un outlier y sabemos que es de magnitud, en el caso de forma destacan muchas más curvas 157, 158, 197, 198, 199, 200 y 201.

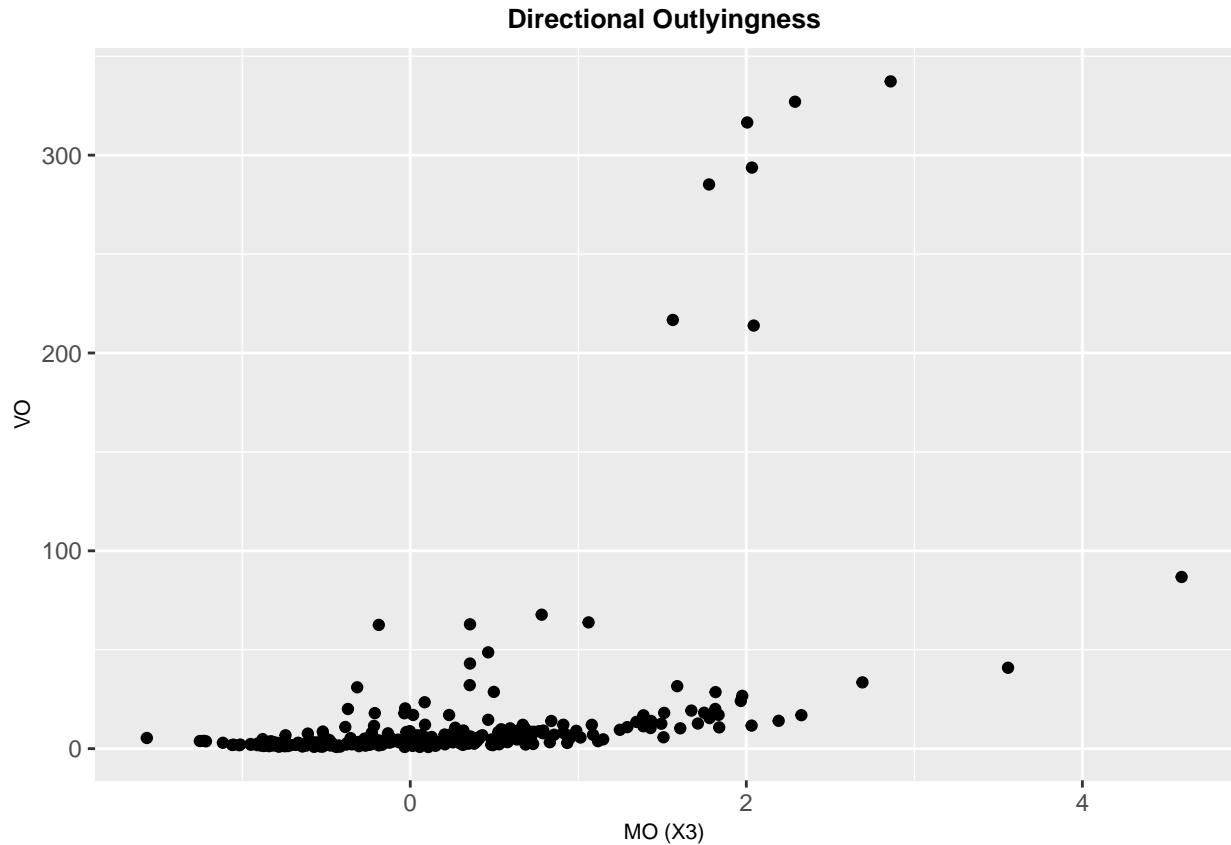


Figure 18:  $MO$  y  $VO$  para el proceso  $X_3$  basado en la profundidad proyectada.

## References

Wenlin Dai and Marc G Genton. Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis*, 131:50–65, 2019.

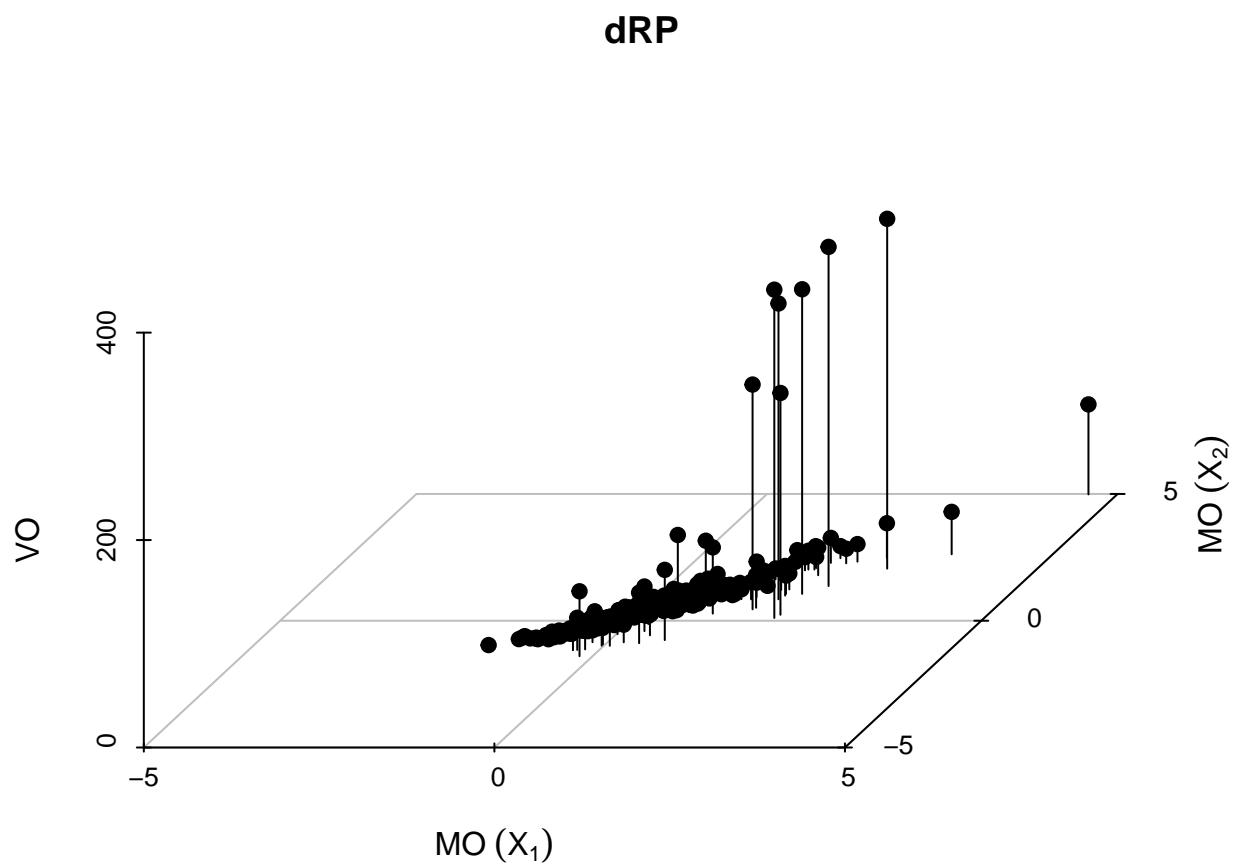


Figure 19: MO y VO bivariado para los procesos  $X_1$  y  $X_2$  basado en la profundidad proyectada.

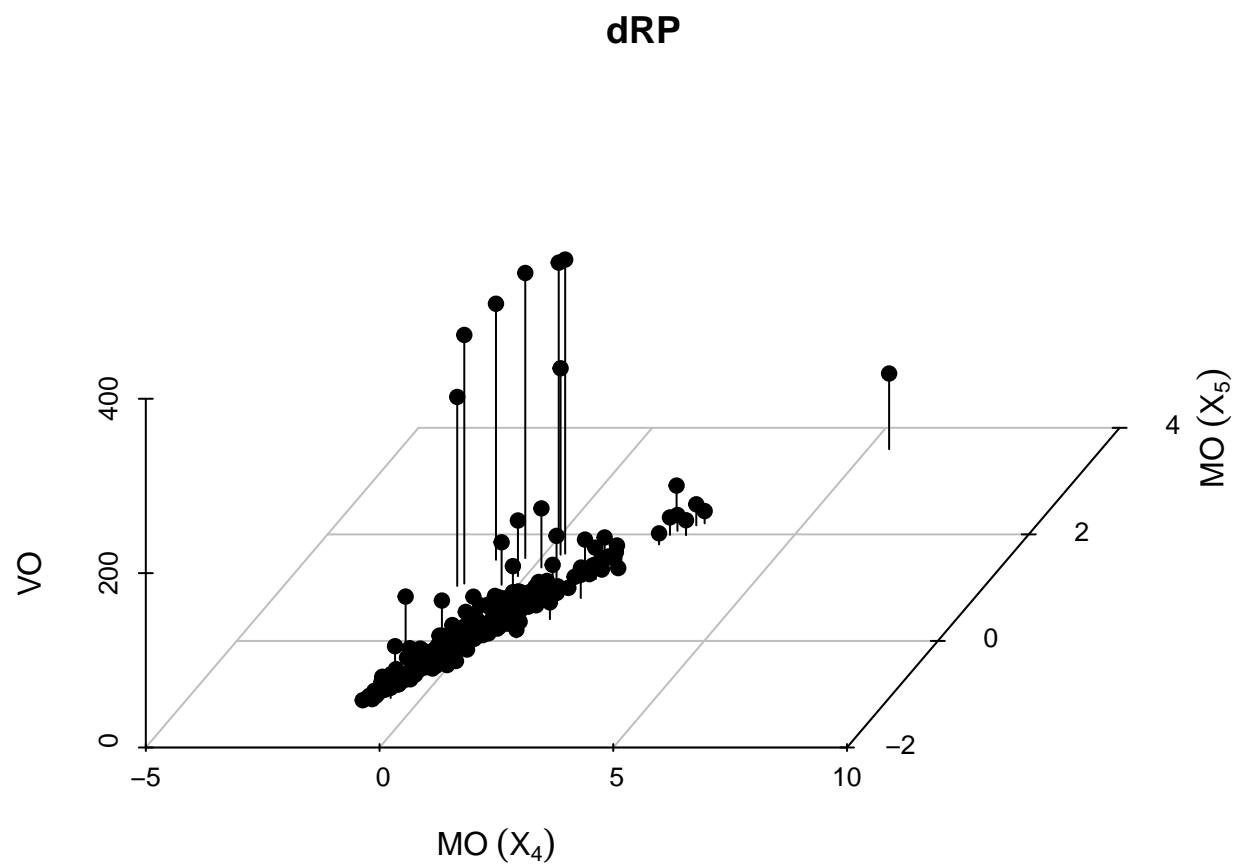


Figure 20: MO y VO bivariado para los procesos  $X_4$  y  $X_5$  basado en la profundidad proyectada.

Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012. URL <http://www.jstatsoft.org/v51/i04/>.

Francesca Ieva, Anna Maria Paganoni, Juan Romo, and Nicholas Tarabelloni. roahd Package: Robust Analysis of High Dimensional Data. *The R Journal*, 11(2):291–307, 2019. doi: 10.32614/RJ-2019-032. URL <https://doi.org/10.32614/RJ-2019-032>.

Sara López-Pintado and Juan Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009.

Naveen N Narisetty and Vijayan N Nair. Extremal depth for functional data and applications. *Journal of the American Statistical Association*, 111(516):1705–1714, 2016.

Oluwasegun Taiwo Ojo, Rosa Elvira Lillo, and Antonio Fernandez Anta. *fdaoutlier: Outlier Detection Tools for Functional Data Analysis*, 2021. URL <https://CRAN.R-project.org/package=fdaoutlier>. R package version 0.2.0.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL <https://www.R-project.org/>.

J. O. Ramsay, Spencer Graves, and Giles Hooker. *fda: Functional Data Analysis*, 2020. URL <https://CRAN.R-project.org/package=fda>. R package version 5.1.9.

James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2007.

Dalia Valencia, Rosa Lillo, Juan Romo, EG Bongiomo, E Salinelli, A Goia, and P Vieu. Spearman coefficient for functions. *Contributions in Infinitesimal Statistics and Related Topics*, Ed. Bongiomo, EG, Salinelli, E., Goia, A., Vieu, P., Societa Editrice Esculapio, pages 269–272, 2014.

Dalia Valencia, Rosa E Lillo, and Juan Romo. A kendall correlation coefficient between functional data. *Advances in Data Analysis and Classification*, 13(4):1083–1103, 2019.