



LICENCIATURA EN CIENCIAS DE DATOS
Propuesta de tema de tesis

Modelos de regresión funcional: aplicaciones, inferencia y algoritmos para su cómputo

En muchas aplicaciones, los datos observados provienen de fenómenos que pueden ser supuestos como curvas o funciones suaves más que como vectores finitos. El análisis de datos funcionales ha recibido notable atención por su versatilidad y aplicabilidad en estos contextos. Aunque el modelo lineal funcional de respuesta escalar es una extensión natural del modelo lineal usual, puede resultar restrictivo para ciertas aplicaciones prácticas, y por eso se han desarrollado propuestas polinómicas para estos contextos. Por otro lado, los procedimientos estadísticos clásicos son muy sensibles al incumplimiento de las hipótesis que los generaron (errores normales u observaciones equidistribuidas, entre otras). Los métodos robustos, en cambio, tienen como objetivo permitir inferencias válidas cuando el modelo no se cumple exactamente y, al mismo tiempo, ser altamente eficientes bajo el modelo central. Por último, muchas de las rutinas que exigen los procedimientos funcionales han sido desarrolladas en librerías de python y R, aunque R concentra el mayor desarrollo del enfoque robusto. Este contexto permite tener perspectiva de algunos trabajos de tesis que podrían desarrollarse.

Palabras clave: Datos funcionales, Modelos de regresión funcional, Estimación robusta, Implementación de rutinas de análisis de datos funcionales en R

Conocimientos deseables

análisis de datos funcionales, estadística robusta, modelos lineales y modelo lineal generalizado, programación en python y en R

¿Qué podría aprender quien realice esta tesis?

En la elaboración de la tesis se aprenderán conocimientos básicos sobre técnicas de estimación e inferencia en el contexto del análisis de datos funcionales. Dependiendo el tema elegido también se aprenderán nociones de estadística robusta y de desarrollo e implementación de rutinas en R.

Dirección de la tesis

*Parada, Daniela
Instituto de Cálculo y CONICET*

Contacto: dparada@dm.uba.ar

Más información en el pdf a continuación.

Modelos de regresión funcional: aplicaciones, inferencia y algoritmos para su cómputo

DANIELA PARADA¹

¹ *Instituto de Cálculo, CONICET y UBA*

daniela.parada@ic.fcen.uba.ar

Palabras Clave: *Datos funcionales, Modelos de regresión funcional, Estimación robusta, Implementación de rutinas de análisis de datos funcionales en R*

Contexto

En muchas aplicaciones, los datos observados provienen de fenómenos que son continuos en el tiempo o en el espacio y que pueden ser supuestos como curvas o funciones suaves más que como vectores de dimensión finita. El **análisis de datos funcionales** ha recibido considerable atención por su versatilidad y aplicabilidad en estos contextos.

Entre otros, podemos mencionar los libros de [Bosq \(1991, 2000\)](#), [Ramsay y Dalzell \(1991\)](#); [Ramsay y Silverman \(2005, 2002\)](#), [Ferraty y Vieu \(2006\)](#), [Ferraty y Romain \(2011\)](#), [Ferraty *et al.* \(2011\)](#), [Horváth y Kokoszka \(2012\)](#); [Kokoszka y Reimherr \(2017\)](#), [Hsing y Eubank \(2015\)](#), [Aneiros-Pérez *et al.* \(2017\)](#) y los trabajos de [Cuevas \(2014\)](#) y [Goia y Vieu \(2016\)](#), que dan enfoques recientes sobre el tema.

Por otro lado, los modelos de regresión clásicos son una herramienta indiscutida de la estadística y aunque es natural querer buscar una extensión al contexto funcional, el problema intrínseco de la dimensión es algo que no puede eludirse. Autores como [Bosq \(1991\)](#); [Ramsay y Dalzell \(1991\)](#); [Hastie y Mallows \(1993\)](#); [Cardot *et al.* \(1999\)](#), entre otros, han considerado estrategias funcionales para aprovechar la naturaleza subyacente de ciertos fenómenos aplicados. Aspectos generales han sido resumidos en las obras de [Ramsay y Silverman \(2005, 2002\)](#); [Bosq \(2000\)](#); [Hsing y Eubank \(2015\)](#); [Ferraty y Vieu \(2006\)](#). Además, una introducción a la inferencia clásica en el contexto funcional puede hallarse en la obra de [Horváth y Kokoszka \(2012\)](#).

Sin embargo, y como es de esperar, la imposición del modelo lineal puede ser muy restrictiva en algunas aplicaciones prácticas. En ese sentido, [Yao y Müller \(2010\)](#) proponen una extensión de este modelo que incluye una relación funcional polinomial, análoga a la extensión de la regresión lineal simple al caso de la regresión polinomial. [Horváth y Reeder \(2013\)](#) proponen un test de significación para el término cuadrático, lo que permite decidir en este contexto si un enfoque lineal es suficiente o es razonable considerar un operador cuadrático para el modelo.

Por otro lado, los procedimientos estadísticos clásicos son muy sensibles al incumplimiento de las hipótesis que los generaron, como errores normales u observaciones equidistribuidas, entre otras. Los métodos estadísticos robustos, en cambio, tienen como objetivo permitir inferencias válidas cuando el modelo no se cumple exactamente y, al mismo tiempo, ser altamente eficientes bajo el modelo central. Las obras de [Huber \(1981\)](#) y [Yohai \(1987\)](#) dan inicio a las investigaciones en este área, y aspectos generales pueden ser consultados en [Maronna *et al.* \(2019\)](#).

En relación con este contexto, una propuesta de estimadores robustos para el modelo funcional cuadrático puede encontrarse en [Boente y Parada \(2023\)](#), así como una aplica-

ción práctica a datos espectrométricos. Tal aplicación, podría extenderse a un contexto de clasificación, por ejemplo, como el tratado en [Zhu \(2017\)](#), pero en su versión robusta. Más detalles sobre el enfoque clásico de clasificación funcional pueden hallarse a nivel introductorio en [Wang *et al.* \(2016\)](#).

Aunque deseable, el desarrollo de procedimientos robustos no es tarea sencilla, y por ello es frecuente que se utilicen diferentes herramientas visuales de diagnóstico. Así como el boxplot es útil para detectar observaciones atípicas bajo la distribución normal, por ejemplo, el boxplot funcional ([Sun y Genton, 2011](#)) es una herramienta análoga en el contexto de datos funcionales, como curvas. [Genton *et al.* \(2014\)](#) definen los surface boxplots usando una noción de profundidad de volumen, lo que permite ordenar datos que en lugar de ser curvas sean superficies, y dar así una noción equivalente a la del boxplot. Tales boxplots funcionales han sido usados en simulaciones y aplicaciones prácticas como los abordados en los recientes trabajos de [Bianco *et al.* \(2022\)](#) y [Bianco y Boente \(2023\)](#).

Por último, muchas de las rutinas y algoritmos que exigen los procedimientos funcionales han sido desarrolladas en librerías de `python` (`scikit-fda`, por ejemplo) y `R` (como, por ejemplo, `fda`, `fda.usc` y `fds`), aunque muchas de las rutinas de estimación robusta en estadística (no solo en el área datos funcionales) tienen mayor desarrollo en `R` que en `python`.

Propuestas

El contexto anterior permite tener perspectiva de algunos trabajos que podrían desarrollarse en el marco de una tesis para la Licenciatura en Ciencia de Datos de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires. A continuación, resumimos algunas de ellas, que por su diferencia en profundidad y complejidad pueden adaptarse fácilmente a los diferentes intereses de lxs futurxs graduadxs. Las que siguen no son excluyentes ni exhaustivas:

- Inferencia en un conjunto de datos reales de naturaleza funcional bajo el modelo polinómico de regresión. En el caso cuadrático, podría involucrar la aplicación a través de propuestas de inferencia e inferencia robusta.
- Estudio de estimadores, clásicos o robustos, para la clasificación de datos funcionales, incluyendo alguna propuesta de simulación o de aplicación a datos reales.
- Desarrollo de algunas rutinas, clásicas y/o robustas, en una librería de `R`. Tal desarrollo puede ser parcial y no publicarse en `CRAN`, sino ser disponibilizado a través de `GitHub`.
- Implementación de rutinas para el cómputo de los `surface boxplots` en `R`, algo que actualmente solo es posible a través de `scikit-fda` en `python`.
- Estudio de reproducción de resultados de alguno de los trabajos sobre datos funcionales que tratan con datos reales. Esto podría incluir, adicionalmente, la aplicación de algún tipo de test y/o extensión a una propuesta de estimación robusta.
- Evaluación de la *performance* de los modelos de regresión funcional sobre datos reales de alguno de los trabajos citados, en comparación con los resultados que se obtienen vía enfoque *machine learning*.

Referencias

- Aneiros-Pérez, G., Bongiorno, E. G., Cao, R., y Vieu, P. (2017). *Functional Statistics and related fields*. Springer.
- Bianco, A. M. y Boente, G. (2023). Addressing robust estimation in covariate-specific roc curves. *Econometrics and Statistics*.
- Bianco, A. M., Boente, G., y González-Manteiga, W. (2022). Robust consistent estimators for roc curves with covariates. *Electronic Journal of Statistics*, 16(2):4133–4161.
- Boente, G. y Parada, D. (2023). Robust estimation for functional quadratic regression models. *Computational Statistics & Data Analysis*, 187:107798.
- Bosq, D. (1991). Modelization, nonparametric estimation and prediction for continuous time processes. En *Nonparametric Functional Estimation and Related Topics*, pp. 509–529. Springer Science & Business Media.
- Bosq, D. (2000). *Linear processes in function spaces: theory and applications*, volumen 149. Springer Science & Business Media.
- Cardot, H., Ferraty, F., y Sarda, P. (1999). Functional Linear Model. *Statistics & Probability Letters*, 45(1):11–22.
- Cuevas, A. (2014). A partial overview of the Theory of Statistics with Functional Data. *Journal of Statistical Planning and Inference*, 147:1–23.
- Ferraty, F., Goia, A., Salinelli, E., y Vieu, P. (2011). Recent advances on Functional Additive Regression. *Recent Advances in Functional Data Analysis and Related Topics*, pp. 97–102.
- Ferraty, F. y Romain, Y. (2011). *The Oxford Handbook of Functional Data Analysis*. Oxford University Press.
- Ferraty, F. y Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media.
- Genton, M. G., Johnson, C., Potter, K., Stenchikov, G., y Sun, Y. (2014). Surface boxplots. *Stat*, 3(1):1–11.
- Goia, A. y Vieu, P. (2016). An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis*, 146:1–6.
- Hastie, T. y Mallows, C. (1993). [A statistical view of some chemometrics regression tools]: Discussion. *Technometrics*, 35(2):140–143.
- Horváth, L. y Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer Science & Business Media.
- Horváth, L. y Reeder, R. (2013). A test of significance in functional quadratic regression. *Bernoulli*, 19(5A):2120–2151.
- Hsing, T. y Eubank, R. (2015). *Theoretical foundations of Functional Data Analysis with an introduction to Linear Operators*, volumen 997. John Wiley & Sons.

- Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons.
- Kokoszka, P. y Reimherr, M. (2017). *Introduction to Functional Data Analysis*. CRC Press.
- Maronna, R., Martin, D., Yohai, V., y Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons.
- Ramsay, J. y Dalzell, C. J. (1991). Some tools for Functional Data Analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):539–561.
- Ramsay, J. y Silverman, B. (2002). *Applied Functional Data Analysis. Methods and Case Studies*, volumen 77. Springer.
- Ramsay, J. y Silverman, B. (2005). *Functional Data Analysis, 2nd edition*. Series in Statistics. Springer.
- Sun, Y. y Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334.
- Wang, J.-L., Chiou, J.-M., y Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and its application*, 3:257–295.
- Yao, F. y Müller, H. G. (2010). Functional quadratic regression. *Biometrika*, 97(1):49–64.
- Yohai, V. (1987). High breakdown–point and high efficiency robust estimates for regression. *Annals of Statistics*, 15(2):642–656.
- Zhu, Y. (2017). Functional data analysis of spectroscopic data with application to classification of colon polyps. *American Journal of Analytical Chemistry*, 8(4):294–305.