Estimadores robustos en modelos de regresión funcional cuadrática

Graciela Boente

Universidad de Buenos Aires and CONICET

Abstract

El análisis de datos functionales tiene como objetivo proporcionar herramientas para analizar datos recopilados en forma de funciones o curvas que aparecen en campos como la quimiometría, el reconocimiento de imágenes y la espectroscopía, entre otros. Los datos funcionales son intrínsecamente de dimensión infinita y, como se menciona por ejemplo en Wand et al. (2016), esta estructura de dimensión infinita es de hecho una fuente de información. Por esa razón, incluso cuando se registran en una cuadrícula finita de puntos, las observaciones funcionales deben considerarse como elementos aleatorios de algún espacio funcional en lugar de observaciones multivariadas. De esta manera, se pueden resolver algunos de los desafíos teóricos y numéricos que plantea la alta dimensionalidad. Este marco condujo a la extensión de algunos conceptos y modelos clásicos en estadística, como el modelo de regresión lineal o las técnicas de reducción de dimensión del análisis multivariado, al contexto de los datos funcionales, generalmente a través de alguna herramienta de regularización. Una descripción general de las diferentes herramientas para analizar este tipo de datos se puede ver en Ramsay & Silverman (2005) o Horváth & Kokoszka (2012).

El modelo de regresión funcional lineal supone que las observaciones (y_i, X_i), $1 \leq i \leq n$, son i.i.d. y tales que $y_i = \mu + \langle \beta, X_i \rangle + \sigma_0 \, \epsilon_i$, donde $\beta \in L^2(\mathcal{I})$ y $\sigma_0 > 0$ indican los parámetros de regresión y escala, respectivamente. Los errores se suponen independientes de las covariables X_i. Propuestas robustas para este modelo basadas en splines, $B-$splines o $P-$splines fueron considerados en Maronna & Yohai (2013), Boente et al. (2020) y Kalogridis & Van Aelst (2021), mientras que un enfoque basado en direcciones principales robustas fue dada en Kalogridis & Van Aelst (2019).

Sin embargo, el modelo de regresión funcional lineal puede ser demasiado restrictivo en algunas aplicaciones. Como una generalización del modelo de regresión polinomial usual al caso de covariables funcionales, Yao & Müller (2010) extendieron el modelo de regresión funcional lineal al caso donde la dependencia con las covariables es de naturaleza polinomial en los escores en lugar de lineal. En particular, la relación del modelo de regresión funcional cuadrática involucra una función de cuadrado integrable que corresponde al parámetro lineal β y una función bivariada de cuadrado integrable $v(s,t)$ asociada al parámetro cuadrático que corresponde a un operador lineal, auto–adjunto, Hilbert-Schmidt $\Upsilon : L^2(\mathcal{I}) \to L^2(\mathcal{I})$. Las observaciones cumplen un modelo de regresión funcional cuadrática cuando

$$y_i = \alpha + \langle \beta, X_i - \mu \rangle + \langle X_i - \mu, \Upsilon(X_i - \mu) \rangle + \sigma_0 \, \epsilon_i \,,$$

donde $\mu = \mathbb{E}(X_i)$ y $\Upsilon u(t) = \int_{\mathcal{I}} v(s,t) u(s) \, ds$. El término cuadrático $\langle X_i - \mu, \Upsilon(X_i - \mu) \rangle$ que aparece en el modelo refleja que más allá del efecto que los valores $X_i(t)$, $t \in \mathcal{I}$, tienen en la respuesta, los productos $\{X_i(s) X_i(t)\}$, para $s, t \in \mathcal{I}$, se incluyen como predictores adicionales.

Teniendo en cuenta la sensibilidad de estos estimadores cuando existen observaciones atípicas en la muestra, presentaremos una familia de estimadores robustos que combina proyectar las

covariables funcionales sobre estimadores robustos de las primeras direcciones principales y utilizar posteriormente, $MM-$estimadores de regresión lineal. El comportamiento para muestras finitas de la propuesta dada se ilustrará mediante los resultados de un estudio de Monte Carlo y el análisis de un conjunto de datos reales.

Una cuestión importante a tener en cuenta al definir un estimador robusto es si realmente está estimando la cantidad de interés. Esta propiedad se conoce como consistencia en el sentido de Fisher y suele ser un primer paso antes de obtener resultados de consistencia. La Fisher–consistencia se puede describir fácilmente cuando el estimador se puede escribir como un funcional aplicado a la distribución empírica, véase, por ejemplo, Huber & Ronchetti (2009). Se presentarán resultados generales sobre la Fisher–consistencia del funcional asociado, tanto en el caso de covariables que corresponden a procesos de dimensión finita, es decir si $X = \mu + \sum_{k=1}^{p_0} \xi_k \, \phi_k$, donde ϕ_k son las autofunciones del operador de covarianza de X, como para el caso de procesos infinito–dimensionales. Por otra parte, se presentarán algunas propiedades de robustez y consistencia obtenidas para procesos finito–dimensionales. Esta charla corresponde a un trabajo conjunto con Daniela Parada enviado a publicación.

References

Boente, G., Salibián–Barrera, M. & Vena, P. (2020). Robust estimation for semi-functional linear regression models. *Computational Statistics and Data Analysis*, **152**, 107041.

Horváth, L. & Kokoszka, P. (2012). *Inference for Functional Data with Applications*, Springer, New York.

Huber, P. & Ronchetti, E. (2009). *Robust Statistics*. Wiley.

Kalogridis, I. & Van Aelst, S. (2019). Robust functional regression based on principal components. *Journal of Multivariate Analysis*, **173**, 393-415.

Kalogridis, I. & Van Aelst, S. (2021). Robust penalized estimators for functional linear regression. Available at `https://arxiv.org/pdf/1908.08760.pdf`

Maronna, M. & Yohai, V. (2013). Robust functional linear regression based on splines. *Computational Statistics and Data Analysis*, **65**, 46-55.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, Springer, Berlin.

Wang, J. L., Chiou, J. & Müller, H. G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application*, **3**, 257-295.

Yao, F. & Müller, H. G. (2010). Functional quadratic regression. *Biometrika*, **97**, 49-64.