Resumen del trabajo candidato al premio Ramiro Melendreras

Eduardo García Portugués

1. Contexto

El contexto del trabajo candidato se enmarca en los campos de la estadística noparamétrica y la estadística de datos direccionales (*directional statistics*). En estas dos áreas se utilizan métodos de suavizado (*kernel smoothing*) para diseñar nuevas contribuciones a la estimación y a los contrastes de hipótesis sobre la función de regresión $m : \mathbf{x} \in \Omega_q \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \in \mathbb{R}$, siendo $\Omega_q = \{\mathbf{x} \in \mathbb{R}^{q+1} : ||\mathbf{x}|| = 1\}$ la hiperesfera de dimensión q. La aplicación a datos reales está motivada por una representación de la estructura de las proteínas utilizada en el área de la bioinformática.

2. Estado del arte

Los trabajos previos sobre inferencia noparamétrica en la función de regresión m con datos direccionales se han centrado especialmente en la **estimación**. En Wang et al. (2000) se presenta la primera propuesta para estimar m mediante un estimador Nadaraya-Watson (*i.e.*, local constante). Posteriormente, en Di Marzio et al. (2009) se propone un estimador local lineal para el caso específico de datos circulares ($q = 1$), es decir, con $m_1 : \theta \in [-\pi, \pi) \mapsto \mathbb{E}[Y|\Theta = \theta] \in \mathbb{R}$. El argumento clave para esta propuesta es la sustitución en las expansiones de Taylor de los términos $(\Theta_i - \theta)^k$ por $\sin(\Theta_i - \theta)^k$, resultando:

$$m_1(\Theta_i) = m_1(\theta) + m_1'(\theta)\sin(\Theta_i - \theta) + R.$$

Una idea similar se emplea en Di Marzio et al. (2013) para extender el estimador anterior al caso con respuesta circular. Finalmente, en el trabajo (casi coetáneo) de Di Marzio et al. (2014) se propone un estimador local lineal de la función de regresión $m : \mathbf{x} \in \Omega_q \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \in \mathbb{R}$ utilizando expansiones de Taylor del seno y el coseno en la descomposición tangente-normal de un punto en Ω_q. La versión circular de esta propuesta no extiende a la anterior de Di Marzio et al. (2009). En un contexto con mayor complejidad geométrica, Pelletier (2006) propone un estimador Nadaraya-Watson para variedades Riemmanianas y Cheng and Wu (2013) un estimador local lineal que emplea un procedimiento adaptativo de varias etapas.

Con respecto a los **contrastes de hipótesis**, las contribuciones son muy escasas. Hasta la fecha no se ha presentado ningún contraste de bondad de ajuste para funciones de regresión con datos direccionales. Los autores son conscientes de una única propuesta similar: Deschepper et al. (2008), que propone un contraste de significación en el caso particular de la regresión lineal-circular basado en la comparación de medias por sectores. En el contexto diferente de la inferencia sobre la función de densidad existen contrastes de bondad de ajuste recientes (ver Boente et al. (2014) y García-Portugués et al. (2014)).

Desde el punto de vista euclídeo, las referencias más influyentes en este trabajo son: el contraste de bondad de ajuste de Härdle and Mammen (1993) para la función de regresión utilizando un estimador local constante, el libro de Fan and Gijbels (1996) que recoje la teoría de los estimadores polinómico locales y el contraste de bondad de ajuste de Alcalá et al. (1999) para funciones de regresión polinómicas usando el estimador polinómico local.

3. Aportaciones

En este trabajo se propone un estimador local lineal $\hat{m}_{h,p}$ de la función de regresión $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ a partir de una muestra $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n \subset \Omega_q \times \mathbb{R}$. Esta propuesta extiende de forma natural a la de Wang et al. (2000) para el estimador direccional local constante y a la de Di Marzio et al. (2009) para el estimador circular local lineal, algo que no sucede con la propuesta de Di Marzio et al. (2014). Además, la extensión surge partiendo desde una motivación totalmente diferente, en la que se considera un nuevo tipo de expansión de Taylor en el plano tangente a $\mathbf{X}_i$ con el fin de aproximar $m(\mathbf{X}_i)$:

$$m(\mathbf{X}_i) = m(\mathbf{x}) + (\mathbf{B}_\mathbf{x}^T \boldsymbol{\nabla} m(\mathbf{x}))^T (\mathbf{B}_\mathbf{x}^T (\mathbf{X}_i - \mathbf{x})) + o(||\mathbf{X}_i - \mathbf{x}||^2),$$

con $\mathbf{B}_\mathbf{x}$ una matriz semiortonormal $(q+1) \times q$. Esta modificación de la expansión de Taylor usual permite reducir una dimensión en el coeficiente lineal, asegurando que el problema de mínimos cuadrados resultante tenga una solución bien definida. La matriz $\mathbf{B}_\mathbf{x}$ aparece con anterioridad en cálculos del estimador núcleo de la densidad (ver García-Portugués et al. (2013) y García-Portugués et al. (2014)).

La propuesta del estimador $\hat{m}_{h,p}$ permite mantener una complejidad matemática acotada al mismo tiempo que presenta un estimador consistente y flexible, calculable en la práctica y comparable al euclídeo. Para dicho estimador se han obtenido los siguientes resultados teóricos:

1. Sesgo condicional.
2. Varianza condicional.
3. Expresión del núcleo equivalente.
4. Normalidad asintótica puntual.

A partir del estimador noparamétrico $\hat{m}_{h,p}$ se construye un contraste de bondad de ajuste para modelos paramétricos. Este test permite contrastar la hipótesis compuesta $H_0 : m \in \{m_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ (por ejemplo, un modelo lineal: $H_0 : m(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\theta}, \boldsymbol{\theta} \in \Theta$) mediante un estadístico basado en una distancia L^2 ponderada entre el estimador noparamétrico y el estimador paramétrico suavizado bajo H_0:

$$T_n = \int_{\Omega_q} \left(\hat{m}_{h,p}(\mathbf{x}) - \mathcal{L}_{h,p} m_{\hat{\boldsymbol{\theta}}}(\mathbf{x})\right)^2 \hat{f}_h(\mathbf{x}) w(\mathbf{x})\, \omega_q(d\mathbf{x}),$$

donde w es una función peso, $\hat{f}_h(\mathbf{x})$ es el estimador núcleo de la densidad y $\omega_q(d\mathbf{x})$ representa la medida de Lebesgue en Ω_q. Para calibrar el test en la práctica se diseña un procedimiento de remuestreo bootstrap que permite asignar p-valores de forma efectiva y rigurosa.

Para este test se han obtenido los siguientes resultados teóricos:

5. Distribución asintótica.
6. Potencia bajo alternativas locales.
7. Consistencia del remuestreo bootstrap.

Cabe destacar que para la prueba del punto 7 (el resultado más importante en la práctica) se necesita probar 4, que a su vez depende de 3 y 1. Se ha validado empíricamente la convergencia del estadístico a la distribución asintótica mediante un pequeño estudio numérico.

El buen funcionamiento en la práctica del test se analiza mediante un estudio de simulación, en el que se cubren distintos tamaños muestrales, dimensiones y modelos. Por último, la metodología desarrollada sirve para explorar una hipótesis habitualmente asumida en bioinformática: que la longitud de los pseudo-enlaces es en promedio constante (con respecto a los pseudo-ángulos) en la representación C_α de la estructura de una proteína.

4. Idoneidad para el premio

Desde mi punto de vista, considero que el trabajo candidato es meritorio del premio al contener las componentes principales que definen a la estadística moderna:

1. **Resultados teóricos rigurosos** que sientan unas sólidas bases metodológicas: sesgo, varianza y normalidad asintótica del estimador de la regresión; distribución asintótica, potencia bajo alternativas locales y consistencia bootstrap para el contraste de bondad de ajuste.

2. **Procedimientos computacionales** que permiten implementar y comprobar los métodos estadísticos: asignación de p-valores de forma efectiva en la práctica, validación empírica de la distribución asintótica y realización de un completo estudio de simulación.

3. **Aplicaciones a problemas de interés práctico**: estudio de la hipótesis clave en la representación C_α de la estructura de las proteínas empleada en bioinformática y visualización de las regiones donde la desviación de esta hipótesis es mayor.

Referencias

Alcalá, J. T., Cristóbal, J. A., and González-Manteiga, W. (1999). Goodness-of-fit test for linear models based on local polynomials. *Statist. Probab. Lett.*, 42(1):39–46.

Boente, G., Rodríguez, D., and González-Manteiga, W. (2014). Goodness-of-fit test for directional data. *Scand. J. Stat.*, 41(1):259–275.

Cheng, M.-Y. and Wu, H.-T. (2013). Local linear regression on manifolds and its geometric interpretation. *J. Amer. Statist. Assoc.*, 108(504):1421–1434.

Deschepper, E., Thas, O., and Ottoy, J. P. (2008). Tests and diagnostic plots for detecting lack-of-fit for circular-linear regression models. *Biometrics*, 64(3):912–920.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statist. Probab. Lett.*, 79(19):2066–2075.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2013). Non-parametric regression for circular responses. *Scand. J. Stat.*, 40(2):238–255.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2014). Nonparametric regression for spherical data. *J. Amer. Statist. Assoc.*, 109(506):748–763.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

García-Portugués, E., Crujeiras, R. M., and González-Manteiga, W. (2013). Kernel density estimation for directional-linear data. *J. Multivariate Anal.*, 121:152–175.

García-Portugués, E., Crujeiras, R. M., and González-Manteiga, W. (2014). Central limit theorems for directional and linear data with applications. *Statist. Sinica*, to appear.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, 21(4):1926–1947.

Pelletier, B. (2006). Non-parametric regression estimation on closed Riemannian manifolds. *J. Nonparametr. Stat.*, 18(1):57–67.

Wang, X., Zhao, L., and Wu, Y. (2000). Distribution free laws of the iterated logarithm for kernel estimator of regression function based on directional data. *Chinese Ann. Math. Ser. B*, 21(4):489–498.