

1. ARTÍCULOS DE ESTADÍSTICA

ESTADÍSTICA NO PARAMÉTRICA: PASADO, PRESENTE Y FUTURO

Wenceslao González Manteiga y Rosa María Crujeiras Casais

Departamento de Estadística e Investigación Operativa

Universidade de Santiago de Compostela

Resumen

La estadística no paramétrica engloba una serie de técnicas de inferencia cuya característica principal es la ausencia de un modelo paramétrico de distribución subyacente.

Sin pretender revisar todos y cada uno de los conceptos e ideas que conforman este planteamiento, es nuestro objetivo que el lector se familiarice con términos como *técnicas de suavizado*, *Bootstrap*, *verosimilitud empírica* o *datos funcionales*, su porqué, su utilidad, su relevancia en la estadística actual y la posible proyección hacia el futuro.

1. De la distribución empírica a las técnicas de suavizado.

Gran parte de la inferencia estadística paramétrica (Bayesiana o no) trabaja bajo el supuesto de disponer de una muestra aleatoria simple (m.a.s.) $\vec{X} = (X_1, \dots, X_n)$ de una variable poblacional X . A esta variable poblacional se le supone una distribución F_θ con θ un parámetro desconocido finito dimensional. Los diversos métodos de estimación puntual de θ , la construcción de regiones de confianza o los contrastes de hipótesis para dicho parámetro, fueron abordados a lo largo del siglo XX, y resumidos en su parte esencial, en textos como los *Lehmann* ([14] y [15], en sus ediciones revisadas) o los más recientes [24] y [27], entre otros. Estos textos se utilizan además como referencia en cursos de estadística matemática.

En este contexto, la especificación errónea de la distribución poblacional F_θ , o incluso, su total desconocimiento, propiciaron el desarrollo de la inferencia estadística no paramétrica. No se presupone un modelo paramétrico para la distribución poblacional y simplemente se actúa con la filosofía de dejar hablar a los datos. En el enfoque no

paramétrico, la distribución F de la variable poblacional no está caracterizada por un parámetro finito dimensional.

Sin la suposición precisa de un modelo paramétrico de distribución poblacional para la variable de interés, nos encontramos en una encrucijada en la que el problema de inferencia puede ser abordado desde dos vertientes: (a) considerar, como ya se comentó, que el espacio de posibles distribuciones \mathcal{F} no se puede parametrizar de forma finito dimensional, o bien (b) examinar alguna distancia $d(F, F_\theta)$ entre la distribución poblacional desconocida y una posible distribución paramétrica para el modelo poblacional.

La vía (a) dio origen a varios procedimientos de estimación, como los M-estimadores, L-estimadores o R-estimadores. Estos estimadores conforman la extensión de los métodos de estimación de máxima verosimilitud, en el caso de los M-estimadores; de los estadísticos de orden, para los L-estimadores; y de los estadísticos de rango, para los R-estimadores. El objetivo en este caso es hacer inferencia sobre una cierta característica $T(F)$, donde F es la distribución poblacional y T un funcional definido sobre el espacio de distribuciones. Por ejemplo, el libro de Serfling, [22], recoge un resumen didácticamente ordenado del material elaborado en las décadas de los sesenta y setenta. Otros textos coetáneos, como [11] o [13] estudiaron el llamado análisis de la robustez que permitía estudiar la sensibilidad de los diversos estimadores a la desviación de un modelo poblacional.

Por otro lado, los llamados tests de bondad de ajuste nacen del enfoque (b) y han evolucionado desde sus inicios con el más básico Kolmogorov-Smirnov, donde $d(\hat{F}, F_{\hat{\theta}}) = \sup |\hat{F}(x) - F_{\hat{\theta}}(x)|$, \hat{F} es un estimador no paramétrico de la función de dis-

tribución y $\hat{\theta}$ un estimador para θ , a distintos tests alternativos a este (ver, por ejemplo [8]); como los tests del tipo Cramer Von Mises o de tipo χ^2 , por ejemplo. Aunque con diferentes planteamientos, el objetivo es siempre ubicar el modelo de distribución F en una determinada familia paramétrica.

En la mayoría de los procedimientos de estimación o contrastes antes descritos interviene con frecuencia un estimador piloto no paramétrico. En el caso del test de Kolmogorov-Smirnov, este estimador es la la distribución empírica:

$$\hat{F}(x) = F_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq x\}}. \quad (1)$$

Este mismo estimador F_n aparece, por ejemplo, en los métodos de M-estimación, donde un M-estimador es la solución a una ecuación del tipo:

$$\int \Psi(x, t) dF_n(x) = 0, \quad (2)$$

siendo Ψ una función elegida de antemano para el estimador correspondiente.

Una de las principales dificultades en la inferencia no paramétrica surge cuando se pretende estimar cantidades $T(F)$ que dependen de la derivada de F , sin ir más lejos:

$$T(F) = F'(x) = f(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x-h)}{2h}, \quad (3)$$

la función de densidad asociada a la variable poblacional. La estimación de la densidad dio lugar a las denominadas *técnicas de suavizado*. Es así que un estimador natural de (3) es el que viene dado por

$$\hat{f}_h(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}, \quad \text{con } h \rightarrow 0 \quad (4)$$

el llamado estimador de Rosenblatt [21], generalizado por Parzen en [18] a los populares *estimadores tipo núcleo*:

$$\tilde{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (5)$$

donde la función núcleo o *kernel* K es una función de densidad y h se denomina parámetro ventana. El estimador de Rosenblatt se corresponde con la elección de un núcleo uniforme. Una revisión de las

técnicas no paramétricas para la estimación de la densidad puede verse en [23], que representa uno de los primeros textos, de una larga lista, dedicados a las técnicas de suavización publicados en los últimos veinte años.

Las técnicas de suavización no paramétrica, que surgen en el contexto de la estimación de la densidad, rápidamente se extendieron al problema de estimación de otras funciones estadísticas notables, como la función de razón de fallo en fiabilidad $r(x) = \frac{f(x)}{1 - F(x)}$, la función de regresión $m(x) = E(Y|X = x)$ ligada a un vector $(X, Y) \in \mathbb{R}^{q+1}$ ó la varianza condicional $\sigma^2(x) = Var(Y|X = x)$.

Para el caso de la función de regresión, donde la selección de un modelo paramétrico no adecuado puede llevarnos a conclusiones incorrectas, el primer estimador tipo núcleo es el de Nadaraya-Watson (introducido en [16] y [26]), que viene dado por:

$$\tilde{m}(x) = \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)}, \quad (6)$$

donde la concentración local de datos propia del estimador de la densidad (5) se sustituye por el promedio local de la variable respuesta.

También existen otros estimadores tipo núcleo, como el de Priestley-Chao o el de Gasser-Muller (véase [12]). Una excelente revisión de estas técnicas de suavización, tanto para el caso de la densidad como para la regresión, puede verse en [25]. En este texto también se incluyen diversos capítulos dedicados al problema de la selección del parámetro h de suavizado.

Los procedimientos de suavización también se abordaron desde espacios funcionales alternativos al de las funciones derivables.

Si la función de regresión puede escribirse como $m = \sum_{k=1}^{\infty} c_k(m) \phi_k$, donde $\{\phi_k\}_{k=1}^{\infty}$ es un sistema ortogonal completo de \mathcal{L}_2 y $c_k(m)$, $k = 1, 2, \dots$ los coeficientes de Fourier de m , un estimador suavizado natural viene dado por:

$$\widehat{m^O} = \sum_{k=1}^{\hat{M}} \hat{c}_k \phi_k \quad (7)$$

siendo \hat{c}_k los estimadores de los coeficientes de Fourier y \hat{M} el número de coeficientes estimados (ver [2]).

Otra alternativa es la suavización spline $\widehat{m^S}$, donde dicho estimador viene dado por el argumento que minimiza

$$\min_m \frac{1}{n} \sum_{i=1}^n (Y_i - m(X_i))^2 + \lambda \int_0^1 (m^{(p)}(x))^2 dx,$$

y cuya solución es una función del espacio de Sobolev $W_2^p[0, 1]$ (ver [10]). El caso particular de $p = 2$ se corresponden con el conocido *spline cúbico*. También de forma alternativa a los estimadores clásicos tipo núcleo, tuvieron éxito los estimadores de tipo lineal local [6] y más recientemente, de manera alternativa a los estimadores basados en desarrollos ortogonales, los estimadores wavelet [2].

2. El Bootstrap. Todavía actualidad.

En un artículo de *Annals of Statistics* en 1979, [3], Efron introdujo una metodología que alcanzaría su mayor esplendor años después, con el desarrollo computacional: nos referimos al *Bootstrap*.

El Bootstrap es una metodología de inferencia estadística basada en el remuestreo de los datos y cuyo principal inconveniente (aunque no tanto hoy en día) es la necesidad de computación intensiva. Dada una m.a.s. \vec{X} de una variable aleatoria X , denotaremos por \vec{X}^* a las distintas remuestras que se puedan obtener a partir de una estimación de la distribución poblacional. Si estas remuestras se generan a partir de la distribución empírica F_n , el procedimiento se denomina Bootstrap Naive; si la generación de datos se hace a partir de un estimador núcleo

$$\hat{F}_h(x_0) = \int_{-\infty}^{x_0} \hat{f}_h(x) dx$$

tendremos un Bootstrap Suavizado; también existen las alternativas de Bootstrap Paramétrico, Simetrizado, Bootstrap Bayesiano, entre otras opciones.

Consideremos $R(\vec{X}, F) = T(\vec{X}) - T(F)$, siendo $T(\vec{X})$ un estimador de $T(F)$. Si se tiene como objetivo estimar la distribución de $R(\vec{X}, F)$ dada por $G(t) = \mathbb{P}_F\{R(\vec{X}, F) \leq t\}$, el estimador natural

Bootstrap es $\hat{G}^{Boot} = \mathbb{P}_{\hat{F}}\{R(\vec{X}^*, \hat{F}) \leq t\}$. El estimador Bootstrap \hat{G}^{Boot} raramente tiene expresión explícita y debe ser aproximado por Monte Carlo:

$$\widehat{\hat{G}^{Boot}}(t) = \frac{1}{B} \sum_{j=1}^B \mathbf{1}_{\{R(\vec{X}^{*j}, \hat{F}) \leq t\}}, \quad (8)$$

siendo \vec{X}^{*j} , con $j = 1, \dots, B$, remuestras artificiales de la distribución \hat{F} , estimador de F .

La elección que hemos hecho de R es la más sencilla y permitiría inferir sobre la distribución de $T(\vec{X})$ como estimador de $T(F)$, pudiendo así construirse regiones de confianza o plantear contrastes de hipótesis. La adaptación de la metodología Bootstrap a otros problemas notables, como pueden ser el cálculo de la probabilidad de clasificación incorrecta en el análisis discriminante o la construcción de bandas de confianza para una función de distribución, nos llevarían a distintas elecciones de R , que conforman la literatura dedicada a este método en las dos últimas décadas.

Igual que el Barón de Munchausen salió del lago tirando de los cordones de sus botas (en inglés *bootstraps*), del mismo modo Efron presenta una metodología que se apoya en los datos de la muestra observada, para sobre esas observaciones, construir nuevas muestras.

En los últimos quince años han aparecido diversas publicaciones dedicadas a la metodología Bootstrap, como [1] y [5], aunque ya más recientemente, el Bootstrap representa uno o varios capítulos de los textos sobre estadística matemática ([24] o [28]) en los cursos de licenciatura y doctorado.

3. La verosimilitud empírica: la verosimilitud de finales del siglo XX.

Volviendo al planteamiento inicial de considerar una m.a.s. \vec{X} de una variable aleatoria X con distribución F_θ y densidad o masa de probabilidad asociada f_θ , si planteamos la función de verosimilitud $\psi(\theta) = \prod_{i=1}^n f_\theta(X_i)$, el estimador máximo verosímil $\hat{\theta}$ para el parámetro θ , se obtiene como el argumento que maximiza $\psi(\theta)$. Este método fue estudiado a lo largo del siglo pasado y el principio de verosimilitud sigue presente en textos actuales,

como [19], entre otros muchos.

Nuevamente nos encontramos ante el supuesto de una distribución paramétrica para la variable de interés y una vez más, desde el enfoque no paramétrico, se aborda este planteamiento de máxima verosimilitud sin presuponer un modelo poblacional para X . De este modo, cuando no se sabe nada sobre el modelo poblacional, uno puede considerar una verosimilitud de tipo no paramétrico: $\Psi(F) = \prod_{i=1}^n (F(X_i) - F(X_i^-)) = \prod_{i=1}^n \omega_i$, $\vec{\omega} = (\omega_1, \dots, \omega_n)$. De forma natural, se deduce que el estimador de máxima verosimilitud no paramétrico se alcanza en $F = F_n$, dada por (1).

La idea de la *verosimilitud empírica* (o verosimilitud no paramétrica) es reciente, introducida por Owen a finales de los ochenta. Gran parte de sus desarrollos en la década de los noventa se recogen en [17]. La clave del éxito de esta técnica y su creciente desarrollo en los últimos años radica en su flexibilidad y versatilidad para incorporar en la verosimilitud las restricciones que impone el modelo. Así, como ejemplos notables:

- 1) Si sabemos que $\mu_0 = \int x dF(x)$ es la media (conocida) de X , el estimador máximo verosímil para F bajo esta restricción, se obtendría como el argumento que maximiza:

$$\max_{\vec{\omega}} \left\{ \prod_{i=1}^n \omega_i : \sum \omega_i X_i = \mu_0, \omega_i \geq 0, \sum \omega_i = 1 \right\}.$$

- 2) Si suponemos además que la varianza de X es $\sigma_0^2 = \int (x - \mu_0)^2 dF(x)$; el estimador de máxima verosimilitud vendría dado por:

$$\max_{\vec{\omega}} \left\{ \prod_{i=1}^n \omega_i : \sum \omega_i X_i = \mu_0, \sum \omega_i (X_i - \mu_0)^2 = \sigma^2, \omega_i \geq 0, \sum \omega_i = 1 \right\}.$$

De la misma forma que, para el contraste de $H_0 : \theta \in \Theta_0$ vs. $H_a : \theta \notin \Theta_0$, con $\Theta_0 \subset \Theta$, ligado a la clásica razón de verosimilitudes

$$\lambda = \frac{\sup_{\theta \in \Theta_0} \prod_{i=1}^n f_{\theta}(X_i)}{\sup_{\theta \in \Theta} \prod_{i=1}^n f_{\theta}(X_i)} \quad (9)$$

tenemos el test de razón de verosimilitudes, con región crítica $\{-2 \log \lambda > c\}$, lo mismo se puede considerar para la verosimilitud empírica. Supongamos

que se pretende contrastar $H_0 : \mu_0 = \int x dF(x)$. El test de razón de verosimilitudes no paramétrico vendría dado por:

$$R = \frac{\max\{\prod_{i=1}^n \omega_i : \sum \omega_i X_i = \mu_0, \omega_i \geq 0, \sum \omega_i = 1\}}{\max\{\prod_{i=1}^n \omega_i : \omega_i \geq 0, \sum \omega_i = 1\}}$$

$$= \max \left\{ \prod_{i=1}^n \omega_i : \sum \omega_i X_i = \mu_0, \omega_i \geq 0, \sum \omega_i = 1 \right\}$$

Igual que para la razón de verosimilitudes paramétrica, resultados del tipo $-2 \log R \rightarrow^d \chi_m^2$ también se obtienen en este contexto, y en el caso particular anterior resulta $m = 1$.

En el reciente libro de Owen antes citado, [17], se presenta la metodología de verosimilitud empírica en su conjunto para el contexto de regresión, para datos incompletos o con censura, entre otras situaciones notables de la inferencia estadística.

4. La estadística con datos funcionales. La era de los datos de alta dimensión.

La explosión del llamado *data mining* (también conocido como KDD: Knowledge-Discovery in Databases), está produciendo una variedad de situaciones, en diferentes contextos, en que los datos medidos son realmente curvas. Por ejemplo, el nivel de ozono medido durante un día o un indicador de alta frecuencia de un activo financiero en la bolsa a lo largo de una hora. De esta forma, el dato observado es ahora $\chi = \{X(t) : t \in (t_{min}, t_{max})\}$ y podríamos considerar una muestra de *datos funcionales* χ_1, \dots, χ_n que tuviese la misma distribución que χ .

De modo más genérico, una variable aleatoria χ se dice que es una variable funcional si toma valores en un espacio de dimensión infinita. Con este tipo de planteamiento entramos en la era actual y futura de la *estadística no paramétrica con datos funcionales*. Por ejemplo, supongamos que el soporte de la variable funcional es $L^2(0, 1)$ y consideremos un modelo de regresión del tipo: $Y = \mathbf{m}(\mathcal{X}) + \varepsilon$, en el que $Y \in \mathbb{R}$ y ε es una v.a. real de media cero. El objetivo podría ser estimar no paramétricamente el funcional $\mathbf{m}(\cdot) : L^2(0, 1) \rightarrow \mathbb{R}$.

Las técnicas estadísticas para datos reales o vectoriales no se extienden de forma trivial al con-

texto de datos funcionales. Si tenemos un modelo de regresión no paramétrico $Y = m(X) + \varepsilon$ y consideramos un estimador no paramétrico del tipo Nadaraya-Watson, como en (6) nos puede interesar contrastar la hipótesis nula de que el modelo es lineal $m(X) = \beta^T X$. Esto se podría hacer a través de cualquier estadístico del tipo $d(\hat{m}(X), \hat{\beta}^T X)$, que establece una distancia entre el estimador no paramétrico y el paramétrico, por ejemplo, una distancia de tipo L_2 (ver [9]).

En el supuesto de datos funcionales el modelo lineal es ahora del tipo $\mathbf{m}(\mathcal{X}) = \int_0^1 \beta(t) \mathcal{X}(t) dt$ y por tanto también de dimensión infinita, ya que lo desconocido β es también una función. Podemos construir un estimador no paramétrico del funcional \mathbf{m} , análogo al de Nadaraya-Watson del contexto de regresión, que vendría dado por:

$$\hat{\mathbf{m}}(\mathcal{X}) = \frac{\sum_{i=1}^n K\left(\frac{d(\mathcal{X}, \mathcal{X}_i)}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{d(\mathcal{X}, \mathcal{X}_i)}{h}\right)} \quad (10)$$

donde d es una semimétrica asociada al espacio de funciones y K es una función núcleo asimétrica.

En un contexto finito dimensional, en la distribución del estadístico $d(\hat{m}(X), \hat{\beta}^T X)$, la estimación paramétrica puede tener menos peso que la de \hat{m} , ya que sus tasas de convergencia hacia la función teórica son más rápidas. Pero ¿qué ocurre si la distancia se mide en el contexto de datos funcionales? Ambos parámetros, \mathbf{m} y β son de dimensión infinita. Problemas similares a este merecen sin duda el interés de la futura metodología estadística.

Siguiendo esta línea de extensión de las técnicas de regresión no paramétrica a los datos funcionales, [7] es una buena referencia. Un planteamiento alternativo para el estudio de datos funcionales, muy próximo a la metodología de la estimación spline, se recoge en [20].

5. Algunas aplicaciones.

Sin tratar de ser exhaustivos, finalizamos esta nota comentando algunas aplicaciones donde la inferencia no paramétrica juega un papel importante.

Para aproximar la fórmula de Black-Scholes al

comportamiento del mercado, se hace necesario modelizar la volatilidad de manera heterocedástica y por tanto, se requiere la estimación no paramétrica de la varianza.

En los modelos de predicción kriging de la estadística espacial, la suposición de estacionariedad puede resultar restrictiva para llevar a cabo predicciones en problemas de medio-ambiente. Los modelos de predicción han de adaptarse de forma local, y esta adaptación requiere de técnicas de suavización no paramétricas.

Dentro de la estimación de conjuntos, puede plantearse como objetivo estimar el soporte de una cierta variable X , con distribución F desconocida. Una estimación del soporte vendría dada por aquellos puntos donde una estimación no paramétrica de la densidad tomase valores positivos. Este planteamiento podría tener aplicación directa en el reconocimiento de imágenes, entre otras situaciones.

En estos ejemplos, junto con los ya comentados en el apartado dedicado a la estadística con datos funcionales, ¿tendría sentido el suponer un modelo paramétrico? La respuesta no es evidente. Pero lo que sí queda claro es que el planteamiento no paramétrico puede aportar una solución.

6. Conclusiones.

Como se deduce de este recorrido por la inferencia no paramétrica, las razones del porqué de este modo de hacer inferencia son claras. La metodología no paramétrica es, ante todo, flexible, capaz de modelizar sin suposiciones paramétricas restrictivas y sobre todo, rica y elegante desde el punto de vista matemático. Es esta flexibilidad, junto con el desarrollo computacional, lo que ha permitido el éxito de muchas de estas técnicas en los últimos años. Como ejemplo más claro, no sólo la multitud de artículos y referencias en la literatura, sino también sus numerosas aplicaciones en modelos de interés.

Efron, [4], hace una reflexión sobre los dos modos de hacer estadística, todavía enfrentados. La estadística bayesiana dominó el siglo XIX, mientras que la estadística del siglo XX fue más frecuentista. ¿Hacia dónde va la estadística del siglo XXI?

Como Efron sugiere, los nuevos problemas a los que se enfrenta la estadística, sólo podrán ser resueltos combinando ambas filosofías. Desde nuestra posición no creemos osado afirmar que uno de los nexos de unión entre estos enfoques podría ser la estadística no paramétrica.

Referencias

- [1] Davison, A. C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.
- [2] Efromovich, S. (1999). *Nonparametric Curve Estimation. Methods, Theory and Applications*, Springer.
- [3] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1-26.
- [4] Efron, B. (2004). Bayesians, frequentists and scientists. *Text of the 164th ASA presidential address*.
- [5] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman Hall.
- [6] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*, Chapman Hall.
- [7] Ferraty, F. and Vieu, P. (2006). *Nonparametric Modelling for Functional Data*. Springer.
- [8] Gibbons, J.D. and Chakraborti, S. (1992). *Nonparametric Statistical Inference. Third Edition, Revised and Extended*, Marcel Dekker.
- [9] González-Manteiga, W. and Cao, R. (1993). Testing the hypothesis of a general linear model using non parametric regression estimation. *Test*, **2**, 161-188.
- [10] Gu, C. (2002). *Smoothing Spline Anova Models*, Springer.
- [11] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: the Approach Based on Influence Functions*, Wiley.
- [12] Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge.
- [13] Huber, P.J. (1981). *Robust Statistics*, Wiley.
- [14] Lehmann, E.L. (1991). *Theory of Point Estimation*, The Wadsworth & Brooks.
- [15] Lehmann, E.L. (1991). *Testing Statistical Hypothesis*, The Wadsworth & Brooks (2nd edition).
- [16] Nadaraya, E.A. (1964). On estimating regression. *Theor. Prob. Appl.*, **10**, 186-190.
- [17] Owen, A.B. (2001). *Empirical Likelihood*. Chapman & Hall.
- [18] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Stat.*, **33**, 1065-1076.
- [19] Pawtan, Y. (2001). *In all Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford Science Publications.
- [20] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. Springer, 2nd edition.
- [21] Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *Ann. Math. Stat.*, **27**, 832-837.
- [22] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley.
- [23] Silverman, B. (1985). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall.
- [24] Shao, J. (2003). *Mathematical Statistics*, Springer (2nd edition).
- [25] Wand, M. P. and Jones, M.C. (1995). *Kernel Smoothing*, Chapman & Hall.
- [26] Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Ser. A*, **26**, 101-116.
- [27] Wasserman, L. (2004). *All of Statistics. A Concise Course in Statistical Inference*, Springer.
- [28] Wasserman, L. (2006). *All of Nonparametric Statistics*, Springer.