Galician Seminar of
Nonparametric Statistical Inference

Santiago de Compostela
8-9 June 2016

El Departamento de Estadística de Santiago

Orígenes de una investigación

José Antonio Cristóbal
Universidad de Zaragoza

The Santiago Statistics Department

The beginning of a research

José Antonio Cristóbal
University of Zaragoza

December 1978

First professor

Ramiro Melendreras (1944 – 1983)

Luis Coladas Uría (1979). Problemas multiobjetivo: estructuras de dominación. Tesis Doctoral. Universidad de Santiago de Compostela.

Eduardo Ramos Méndez (1979). Mecanismos bioquímicos de acción de los andrógenos sobre la próstata central y las vesículas seminales de la rata. Tesis Doctoral. Universidad de Santiago de Compostela.

PhD's Theses

José Manuel Prada Sánchez (1980). Un concepto de solución para juegos generales con pago multiobjetivo. Tesis Doctoral. Universidad de Santiago de Compostela.

Carmen Carollo Limeres (1981). Nuevos enfoques del concepto de medida de asociación entre variables aleatorias. Distribuciones asintóticas. Tesis Doctoral. Universidad de Santiago de Compostela.

Wenceslao González-Manteiga

Graduated in 1979

Master Dissertation:

"Problemas de Clasificación". Tesina de Licenciatura.
Universidad de Santiago de Compostela. 1980

- Standard Discriminant Analysis
- Discrete variables
- Nonparametric Density Estimation

Thesis

Nonparametric Curves Estimation

Help !

Devroye, L. ⇨ convergences of some nonparametric estimators for density and regression functions

Collomb, G. (1976). Estimation non paramétrique de la régression par la méthode du noyau. PhD thesis, Université Paul Sabatier de Toulouse.

"Construcción axiomática, consistencia y distribuciones asintóticas para estimadores no paramétricos de funciones de densidad y de regresión". Tesis doctoral. Universidad de Santiago de Compostela. Noviembre 1982.

- Methodology and convergences for np-estimators:
  histogram, kernel, splines, interpolations,
  penalizations, nearest neighbors, orthogonal series.

- Axiomatic on a series of delta functions for building general np-density and regression. Convergence properties.

(Previous Communication: Salamanca, april 1982)

From December 1983 to October 1985

- M. Ángeles Fernández Sotelo (1984). "Combinación de métodos bayesianos y no paramétricos para el estudio de la razón de fallo en teoría de la fiabilidad". Tesis Doctoral.

- M. Ángeles Fernández Fernández (1984). "Un estudio dinámico de cuestiones notables en Teoría de Inversión". Tesis Doctoral.

- Pedro Faraldo Roca (1984). "Nuevos aportes de métodos no paramétricos a la teoría de la regresión paramétrica". Tesis Doctoral.

- Domingo Docampo Amoedo (1984). "Decisión en ambiente de incertidumbre parcial". Tesis Doctoral.

1984. Ramiro Melendreras Research Award

“Obtención del sesgo, varianza y error cuadrático medio de una familia axiomática para estimadores no paramétricos de la función de densidad y de regresión”. XIV Reunión de la Sociedad de Estadística e Investigación Operativa. Granada, 1984.

Beginning the collaboration with W. Stute

“Obtención de la eficiencia de una nueva clase de estimadores de los parámetros de un modelo de regresión lineal mediante métodos no paramétricos”. III Meeting International in the Basque Country. Universidad de Lejona. Bilbao. 1985.

“Empirical Processes”, By W. Stute (University of Giessen, Germany). Departamento de Estadística. Universidad de Santiago de Compostela. 1986.

First PhD's Theses directed by Wences

- Juan Manuel Vilar Fernández (1987). “Estimación no paramétrica de la función de densidad y de predicción en series de tiempo”. Universidad de Santiago de Compostela.

- M. Carmen Cadarso Suárez (1990). “Nuevos aportes en la estimación no paramétrica y paramétrica de la regresión con datos censurados”. Universidad de Santiago de Compostela.

- Ricardo Cao Abad (1990). “Aplicaciones y nuevos resultados del método Bootstrap en la estimación no paramétrica de curvas”. Universidad de Santiago de Compostela.

Our Collaborations

Cristóbal, J.A., Faraldo, P. and González-Manteiga, W. (1987). “*A class of linear regression parameter estimators constructed by nonparametric estimation*”. Ann Stat.

Proyecto: “Estimación no paramétrica de curvas: Análisis de problemas notables”. Años 1992-95.

Alcalá, J.T., Cristóbal, J.A. and González-Manteiga, W. (1999). “Goodness-of-fit test for linear models based on local polynomials. Statist. Probabb. Lett.

Ojeda, J.L., González-Mantgeiga, W., Cristóbal, J.A. (2015) “Testing regression models with selection-biased data” Ann. Inst. Stat. Math.

Biased sampling

Also, let X^w be the observed (recorded or intercepted) inter-arrival time covering the fixed time point t_0.
It has density function ("length-biased density function")

$$f_w(x) = \frac{xf(x)}{\mu}, x > 0$$

where f is the density of the inter-arrival times, X.

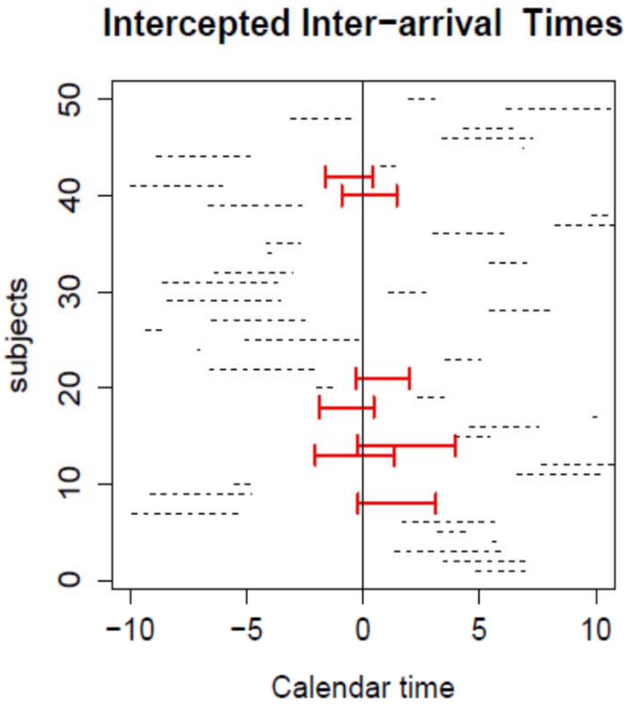

Figure : Inter-arrival times for intercepted subjects at $t_0 = 0$

The greater the length of an interval between two consecutive occurrences, the greater the probability that this will be cut by t_0 and, therefore, included in our sample

Generalizations - Weighted distributions:

When the chance of including an observation x is w(x), the pdf of the recorded random variable X^w is:

$$f^{w}(x)=\frac{w(x)f(x)}{E[X(x)]}$$

The original distribution is not reproduced, although we will usually need to make inferences about it.

The literature contains a number of important examples involving weighted distributions, in a wide range of very diverse contexts: econometrics, marketing, survival analysis, biomedicine and physics among others.

Patil, G. and Rao, C.R. (1978). “Weighted distributions”. In Encyclopedia of Statistical Sciences. John Wiley.

Construction of nonparametric maximum likelihood estimators (NPMLE) of the original distribution function with this type of data:

- Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. Ann. Statist.
- Vardi, Y. (1985). Empirical distributions in selection bias models. Ann. Statist.

Kernel density estimation:
Jones, M.C. (1991). Kernel density estimation for length biased data. Biometrika.

In a multivariate context:
Ahmad, I.A. (1995). On multivariate kernel estimation for samples from weighted distributions. Statist. Probab. Lett.

Regression estimation. Length-biased response variable.

$$f_{XY}^{w}(x) = \frac{y f_{XY}(x,y)}{\mu_Y}$$

where $\mu_Y = \int y f_{XY}\,(x,y)dxdy < \infty$ is the mean of Y.

If $m(x) = E(Y|x)$ is the regression function:

$$f_{Y}^{w}(y) = \frac{y f_{Y}(y)}{\mu_Y}$$

$$f_{Y|X}^{w}(y|x) = \frac{y f_{Y|X}(y|x)}{m(x)}$$

Such as both the marginal density of Y, as well as the conditioned density Y|X also correspond to length biased distributions. Then:

$$E_w\,(Y|X = x) = m(x)[1 + c^2(x)]$$

With $c^2(x) = \frac{\sigma^2(x)}{m^2(x)} = Var(Y|X = x)/E^2(Y|X = x)$ being the conditional variation coefficient.

Direct application of kernel regression estimators to length biased data produces inconsistencies!!!

Cristóbal, J.A. and Alcalá, J.T. (2000). Nonparametric regression estimators for length biased data. J. Stat. Plann. Inference.

First solution: Given that $g(x) = E_w(Y^{-1}|X = x) = m^{-1}(x)$, the ordinary regression kernel estimators $\hat{g}(x)$ applied to the data (X_i, Y_i^{-1}) are consistent for *g(x)*.

Estimator: $\hat{g}_{lp}^{-1}(x)$, the reciprocal of the local polynomial estimator applied to (X_i, Y_i^{-1})

(a smooth harmonic mean)

Second solution: Based on the NPMLE of the distribution function for length biased data has a mass proportional to Y_i^{-1} at each sampling point (X_i, Y_i).

Estimator: weighted local least squares estimator $\hat{\beta}_0$ including the proportional weight:

$$min \sum_i \{Y_i - \beta_0 - \cdots - \beta_p (X_i - x)^p\}^2 K_h\,(X_i - x) Y_i^{-1}$$

(it can be applied to case with general biased data)

(Backward) Recurrence Times

Let U be an Uniform(0,1) random variable independent of X^w, then

$$Y =_d X^w * U$$

(Multiplicative censoring model)

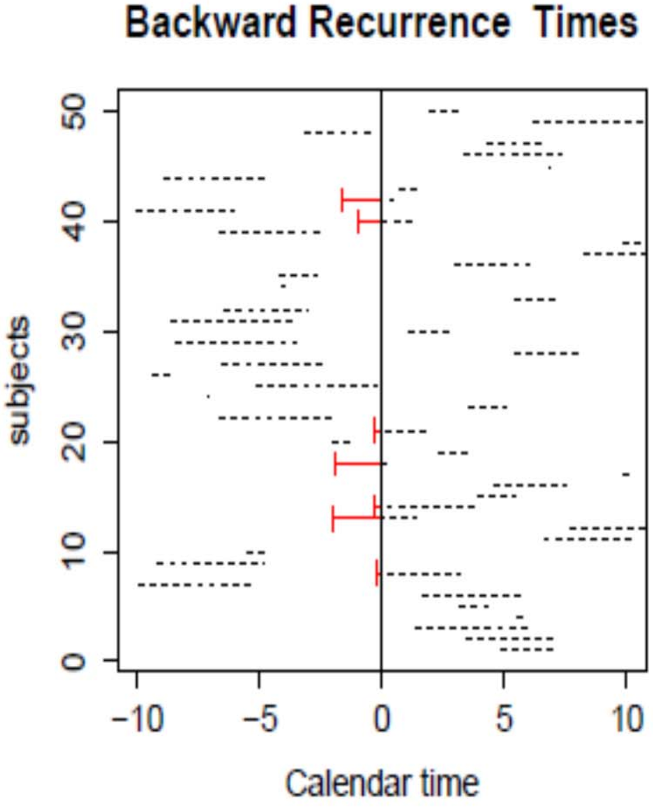

Figure : Backward Recurrence Times as multiplicative censored data

Analysis with a covariate Z:

$$f^{w}_{X|Z}(x|z) = \frac{x f_{X|Z}(x|z)}{E(X|z)}$$

Backward Recurrence Times:

$$f_{Y|Z}(y|z) = \frac{\int_y^\infty f_{X|Z}(x|z)dx}{E(X|z)} = \frac{1 - F_{X|Z}(y|z)}{E(X|z)}$$

When no covariates are present, the NPMLE of the density function is the Grenander estimator: it is computed by the pool-adjacent-violators-algorithm (PAVA)

To avoid the inconsistence at the origin: estimation from a penalized NPMLE problem:
Woodroofe, M. and Sun, J.Y. (1993). “A penalized maximum likelihood estimate of f(0+) when f is non-increasing.”. Stat. Sin.

With a covariate Z, estimation of the regression function:

- Cristóbal, J.A., Alcalá, J.T. and Ojeda, J.L. (2007). “Nonparametric estimation of a regression function from backward recurrence times in a cross-sectional sampling”. Lifetime Data Anal.

General Estimation of the conditional distribution:

- Cristóbal, J.A. and Alcalá, J.T. (2014). “Nonparametric estimation of a conditional distribution for inter-occurrence times, through waiting times in a cross-sectional sampling”. 2nd Conference of the International Society of Nonparametric Statistics. Cádiz.

September 2011. Zaragoza